BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
1 February 2001 (01.02.2001)

PCT

(10) International Publication Number
WO 01/07579 A2

(51) International Patent Classification[7]: C12N 9/00

(21) International Application Number: PCT/US00/20674

(22) International Filing Date: 27 July 2000 (27.07.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/145,898 27 July 1999 (27.07.1999) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/145,898 (CIP)
Filed on 27 July 1999 (27.07.1999)

(71) Applicant *(for all designated States except US)*: THE SALK INSTITUTE FOR BIOLOGICAL STUDIES [US/US]; 10010 North Torrey Pines Road, La Jolla, CA 92037 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: NOEL, Joseph, P. [US/US]; 7280 Park Village Road, San Diego, CA 92129 (US). FERRER, Jean-Luc [FR/FR]; 3, allee Vasco de Balboa, F-38090 Ville Fontaine (FR). JEZ, Joseph [US/US]; 5345 La Jolla Boulevard, La Jolla, CA 92037 (US). AUSTIN, Mike [US/US]; Apartment G, 9192 Regents Road, La Jolla, CA 92037 (US). BOWMAN, Marianne [US/US]; 10034 Riverhead Drive, San Diego, CA 92129-3217 (US).

(74) Agent: REITER, Stephen, E.; Gray Cary Ware & Freidenrich LLP, Suite 1600, 4365 Executive Drive, San Diego, CA 92121 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/07579 A2

(54) Title: METHODS AND COMPOSITIONS FOR DETERMINING ENZYMATIC ACTIVITY

(57) Abstract: The present invention comprises crystalline polyketide synthases, isolated non-native polyketide synthases having the structural coordinates of said crystalline polyketide synthases, and nucleic acids encoding such non-native polyketide synthases. Also disclosed are methods of predicting the activity and/or substrate specificity of putative polyketide synthase, methods of identifying potential polyketide synthase substrates, and methods of identifying potential polyketide synthase inhibitors.

METHODS AND COMPOSITIONS FOR DETERMINING ENZYMATIC ACTIVITY

FIELD OF THE INVENTION

The present invention relates to methods for designing mutant polyketide synthases, and to predicting the activity and/or substrate specificity of putative native and mutant polyketide synthases. The present invention further relates to methods for identifying polyketide synthase substrates and/or inhibitors.

BACKGROUND

Advances in molecular biology have allowed the development of biological agents useful in modulating protein or nucleic acid activity or expression, respectively. Many of these advances are based on identifying the primary sequence of the molecule to be modulated. For example, determining the nucleic acid sequence of DNA or RNA allows the development of antisense or ribozyme molecules. Similarly, identifying the primary sequence allows for the identification of sequences that may be useful in creating monoclonal antibodies. However, often the primary sequence of a protein is insufficient to develop therapeutic or diagnostic molecules due to the secondary, tertiary or quartenary structure of the protein from which the primary sequence is obtained. The process of designing potent and specific inhibitors or activators has improved with the arrival of techniques for determining the three-dimensional structure of an enzyme or polypeptide to be modulated.

The phenylpropanoid synthetic pathway in plants produces a class of compounds know as anthocyanins, which are used for a variety of applications. Anthocyanins are involved in pigmentation and protection against UV photodamage, synthesis of anti-microbial phytoalexins, and are flavonoid inducers of *Rhizobium* modulation genes l-4. As medicinal natural products, the phenylpropanoids exhibit cancer chemopreventive activity, as well as anti-mitotic, estrogenic, anti-malarial, anti-oxidant, and antiasthmatic activities. The benefits of consuming red wine, which contains significant amounts of 3,4',5-trihydroxystilbene (resveratrol) and other phenylpropanoids, highlight the dietary importance of these compounds. Chalcone

synthase (CHS), a polyketide synthase, plays an essential role in the biosynthesis of plant phenylpropanoids.

An improvement in the understanding of the structure/function of these enzymes would allow for the exploitation of the synthetic capabilities of known enzymes for production of useful new chemical compounds, or allow for the creation of novel non-native enzymes having new synthetic capabilities. A need exists, therefore, for a detailed understanding of the molecular basis of the chemical reactions involved in polyketide synthesis. The present invention addresses this and related needs.

SUMMARY OF THE INVENTION

In accordance with the present invention there are presented crystalline polyketide synthases and the three-dimensional coordinates derived therefrom. Three-dimensional coordinates have been obtained for an active form of chalcone synthase and several inactive mutants thereof, both with and without substrate or substrate analog. Similar results have been obtained for the polyketide synthases stilbene synthase and pyrone synthase.

One aspect of the present invention that is made possible by results described herein is that the three-dimensional properties of polyketide synthase proteins are determined, in particular the three-dimensional properties of the active site. The invention features specific coordinates of at least fourteen α carbon atoms defined for the active site in three-dimensional space. R-groups attached to said α-carbons are defined such that mutants can be made by changing at least one R-group found in the synthase active site. Such mutants may have unique and useful properties. Thus, in another embodiment of the invention, there are provided isolated non-native (*e.g.*, mutant) synthase(s) having at least fourteen active site α-carbons having the structural coordinates disclosed herein and one or more R-groups other than those found in native chalcone synthase(s).

The three-dimensional coordinates disclosed herein can be employed in a variety of methods. The polyketide synthase used in the crystallization studies

disclosed herein is a chalcone synthase derived from *Medicago sativa* (alfalfa). A large number of proteins have been isolated and sequenced which have primary amino acid sequence similar to that of chalcone synthase, but for which substrate specificity and/or product is unknown. Thus, in another embodiment of the present invention, there are provided methods for predicting the activity and/or substrate specificity of a putative polyketide synthase. There are further provided methods for identifying potential substrates for a polyketide synthase, as well as inhibitors thereof.

Other aspects, embodiments, advantages, and features of the present invention will become apparent from the following specification.

BRIEF DESCRIPTION OF FIGURES

Figure 1A presents the chemical structures of chalcone, naringenin, resveratrol, and cerulenin. **Figure 1B** presents final SIGMAA-weighted 2Fo-Fc electron density map of the CHS-resveratrol complex in the vicinity of the resveratrol binding site. The map is contoured at 1σ.

Figure 2A shows a ribbon representation of the CHS homodimer. The approximate alpha carbon positions of Met 137 from each of the monomers are labeled accordingly. Naringenin completely fills the coumaroyl-binding and cyclization pockets while the CoA binding tunnels are highlighted by black arrows. Produced with MOLSCRIPT and rendered with POV-Ray. **Figure 2B** presents a stereoview of the monomer's alpha carbon backbone. The orientation of the left-hand monomer is exactly the same as in **Figure 2A**. Every twenty residues are numbered starting with residue 3 and include the C-terminal residue, 389.

Figure 3 shows a comparison of chalcone synthase and 3-ketoacyl-CoA thiolase. Ribbon view of the CHS monomer is oriented perpendicular to the dimer interface. The active site cysteine (Cys 164) and the location of bound CoA are rendered as ball and stick models. In addition, strands β1d and β2d of the cyclization pocket are noted. The reaction catalyzed by CHS is illustrated with the coumaroyl- and malonyl-derived portions of chalcone, respectively. The thiolase monomer is depicted in the same orientation as CHS with the Active site cysteine (Cys 125)

modeled and the reaction of thiolase as indicated. Figure prepared with MOLSCRIPT and rendered with POV-Ray.

Figure 4 collectively shows structures of CHS-Acyl-CoA complexes. The ribbon diagram in panel **Figure 4A** (on the top left) is the same as **Figure 2A**. The CoA binding region depicted in stereo is bounded by a black box in the upper ribbon diagram. Close-up stereoviews of the $C_{164}S$ mutant CoA binding region for the malonyl-and hexanoyl-CoA complexes are depicted in **Figures 4B** and **4C**, respectively. This mutant retains decarboxylation activity and an acetyl-CoA complex is observed crystallographically for the malonyl-CoA complex. In each complex, placement of the Met 137 loop originating from the dyad-related molecule spatially defines one wall of the cyclization pocket. Hydrogen bonds are depicted as spheres. Figure prepared with MOLSCRIPT and rendered with POV-Ray.

Figure 5A shows the CHS-naringenin complex viewed down the CoA-binding tunnel. The ribbon diagram at the top left has been rotated 90 degrees around the y-axis from the orientation shown in **Figure 2A**. This view approximates the global orientation of the CHS dimer used for the close-up view of the natingenin binding site depicted in stereo. Again, the black box highlights the region of CHS shown in stereo close-up. Hydrogen bonds are depicted as dashed cylinders. **Figure 5B** illustrates a comparison of the CHS apoenzyme, CHS-naringenin, and CHS-resveratrol structures. Protein backbone atoms for the three refined structures (apoenzyme, naringenin, and resveratrol) were superimposed by least squares fit in O. The position of bound naringenin and resveratrol are shown. For reference, a modeled low energy conformation of chalcone is indicated by dashed cylinders. Strands β1d and β2d for each complex are also depicted (see **Figure 3**). β2d does not change in all the complexes examined, but β1d moves in the CHS-resveratrol complex. **Figure 5C** presents representative sequence alignment of the β1d -β2d region is given with positions 255, 266, and 268 highlighted. The first three sequences follow a CHS-like cyclization pathway, while the last three use the STS-cyclization pathway. Figure prepared with MOLSCRIPT and rendered with POV-Ray.

Figure 6 presents the proposed reaction mechanism for chalcone synthesis. The three boxed regions labeled 1, 2, and 3 depict the addition of acetate units derived from malonyl-CoA during the elongation of polyketide intermediates. Box I is depicted in expanded fashion to illustrate the mechanistic details governing the decarboxylation, enolization, and condensation phase of ketide elongation. Smaller black arrows depict the flow of electrons. Each acetate unit of the malonyl-CoA thioesters is coded to emphasize the portions of chalcone derived from each of three elongation reactions using malonyl-CoA. Cyclization and aromatization of the enzyme bound tetraketide leads to formation of chalcone. Hydrogen bonds are shown as dashed lines. Coenzyme A is symbolized as a circle.

Figure 7 collectively presents three-dimensional models of the elongation and cyclization reaction in CHS and STS. Views are shown in stereo. **Figure 7A** illustrates the elongation of the triketide covalently attached to Cys 164 by the acetyl-CoA carbanion produces the tetraketide CoA thioester reaction intermediate that subsequently reattaches to Cys 164. **Figure 7B** illustrates the folding of the tetraketide intermediate in CHS positions the oxygen of Cl near the hydrogen of C6 facilitating internal proton transfer and expulsion of chalcone upon cyclization. **Figure 7C** illustrates alternate folding of the tetraketide intermediate and positioning of the oxygen of C7 near the hydrogen of C2 in STS allows formation of resveratrol using an internal proton transfer followed by hydrolysis and decarboxylation. Rendered and dashed lines illustrate potential hydrogen bonding interactions. Figure prepared with MOLSCREPIT and rendered with POV-Ray.

Figure 8 presents a comparison of the active site volumes of CHS and GCHS2. The active site volumes available for binding ketide intermediates were calculated with VOID00 for the CHS-COA complex and for a homology model of GCHS2 with CoA. The cavities are shown as a wire mesh. The homology model of GCHS2 was generated using MODELER and the volume calculated and displayed as for CHS. The numbering scheme is for alfalfa CHS homodimer. Figure prepared with MOLSCRIPT and rendered with POV-Ray.

Figure 9 shows an example of a computer system in block diagram form.

DETAILED DESCRIPTION OF THE INVENTION

The phenylpropanoid synthetic pathway in plants produces a class of compounds know as anthocyanins, which are used for a variety of applications. Anthocyanins are involved in pigmentation and protection against UV photodamage, synthesis of anti-microbial phytoalexins, and are flavonoid inducers of *Rhizobium* modulation genes l-4. As medicinal natural products, the phenylpropanoids exhibit cancer chemopreventive activity, as well as anti-mitotic, estrogenic, anti-malarial, anti-oxidant, and antiasthmatic activities. The benefits of consuming red wine, which contains significant amounts of 3,4',5-trihydroxystilbene (resveratrol) and other phenylpropanoids, highlight the dietary importance of these compounds.

Polyketides are a large class of compounds and include a broad range of antibiotics, immunosuppressants and anticancer agents which together account for sales of over $5 billion per year. Polyketides are molecules which are an extremely rich source of bioactivities, including antibiotics (*e.g.*, tetracyclines and erythromycin), anti-cancer agents (*e.g.*, daunomycin), immunosuppressants (*e.g.*, FK506 and rapamycin), and veterinary products (*e.g.*, monensin) and the like. Many polyketides (produced by polyketide synthases) are valuable as therapeutic agents. Polyketide synthases are multifunctional enzymes that catalyze the biosynthesis of a huge variety of carbon chains differing in length and patterns of functionality and cyclization.

Chalcone synthase (CHS), a polyketide synthase, plays an essential role in the biosynthesis of plant phenylpropanoids. CHS supplies 4,2',4',6'-tetrahydroxychalcone (chalcone) to downstream enzymes that synthesize a diverse set of flavonoid phytoalexins and anthocyanin pigments. Synthesis of chalcone by CHS involves the sequential condensation of one p-coumaroyl- and three malonyl-Coenzyme-A (CoA) molecules (Kreuzaler and Hahlbrock, Eur. J. Biochem. 56:205-213, 1975). After initial capture of the p-coumaroyl moiety, each subsequent condensation step begins

with decarboxylation of malonyl-CoA at the CHS active site; the resulting acetyl-CoA carbanion then serves as the nucleophile for chain elongation.

Ultimately, these reactions generate a tetraketide intermediate that cyclizes by a Claisen condensation into a hydroxylated aromatic ring system. This mechanism mirrors those of the fatty acid and polyketide synthases but with significant differences. CHS uses CoA-thioesters for shuttling substrates and intermediate polyketides instead of the acyl carrier proteins used by the fatty acid synthases. Also, unlike these enzymes, which function as either multichain or multimodular enzyme complexes catalyzing distinct reactions at different active sites, CHS functions as a unimodular polyketide synthase and carries out a series of decarboxylation, condensation, cyclization, and aromatization reactions at a single active site.

A number of plant polyketide synthases related to CHS by sequence identity, including stilbene synthase (STS), bibenzyl synthase (BBS), and acridone synthase (ACS), share a common chemical mechanism, but differ from CHS in their substrate specificity and/or in the stereochemistry of the polyketide cyclization reaction. For example, STS condenses one coumaroyl- and three malonyl-CoA molecules, like CHS, but synthesizes resveratrol (resveratrol) through a structurally distinct cyclization intermediate.

While the cloning of nearly 150 CHS-related genes, and characterization of some of these proteins, provides insight into their biological function, it remains unclear how these enzymes perform multiple decarboxylation and condensation reactions and how they dictate the stereochemistry of the final polyketide cyclization reaction. Furthermore, despite significant advances in the biosynthetic manipulation of structurally complex and biologically important natural products, there remains a lack of structural information on polyketide synthases from any source.

As used herein, "naturally occurring amino acid" and "naturally occurring R-group" includes L-isomers of the twenty amino acids naturally occurring in proteins. Naturally occurring amino acids are glycine, alanine, valine, leucine, isoleucine, serine, methionine, threonine, phenylalanine, tyrosine, tryptophan, cysteine, proline,

histidine, aspartic acid, asparagine, glutamic acid, glutamine, arginine, and lysine. Unless specially indicated, all amino acids referred to in this application are in the L-form.

"Unnatural amino acid" and "unnatural R-group" includes amino acids that are not naturally found in proteins. Examples of unnatural amino acids included herein are racemic mixtures of selenocysteine and selenomethionine. In addition, unnatural amino acids include the D or L forms of, for example, nor-leucine, para-nitrophenylalanine, homophenylalanine, para-fluorophenylalanine, 3-amino-2-benzylpropionic acid, homoarginines, D-phenylalanine, and the like.

"R-group" refers to the substituent attached to the α-carbon of an amino acid residue. An R-group is an important determinant of the overall chemical character of an amino acid. There are twenty natural R-groups found in proteins, which make up the twenty naturally occurring amino acids.

"α-carbon" refers to the chiral carbon atom found in an amino acid residue. Typically, four substituents will be covalently bound to said α-carbon including an amine group, a carboxylic acid group, a hydrogen atom, and an R-group.

"Positively charged amino acid" and "positively charged R-group" includes any naturally occurring or unnatural amino acid having a positively charged side chain under normal physiological conditions. Examples of positively charged, naturally occurring amino acids include arginine, lysine, histidine, and the like.

"Negatively charged amino acid" and "negatively charged R-group" includes any naturally occurring or unnatural amino acid having a negatively charged side chain under normal physiological conditions. Examples of negatively charged, naturally occurring amino acids include aspartic acid, glutamic acid, and the like.

"Hydrophobic amino acid" and "hydrophobic R-group" includes any naturally occurring or unnatural amino acid having an uncharged, nonpolar side chain that is relatively insoluble in water. Examples of naturally occurring hydrophobic amino acids are alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, methionine, and the like.

"Hydrophilic amino acid" and "hydrophilic R-group" includes any naturally occurring or unnatural amino acid having a charged polar side chain that is relatively soluble in water. Examples of naturally occurring hydrophilic amino acids include serine, threonine, tyrosine, asparagine, glutamine, cysteine, and the like.

"Mutant" or "mutated synthase" refers to a polyketide synthase polypeptide, having the three-dimensional coordinates as set forth in Protein Data Bank (PDB) Accession No. 1BI5 (the content of which is incorporated herein by reference in its entirety), and having R-groups on each α-carbon other than the prescribed arrangements of R-groups associated with each α-carbon of a known isolated polyketide synthase (Accession No. 1BI5). Examples of mutant or mutated synthase polypeptides include those having Protein Data Base Accession No. 1D6F, 1K6I, and 1D6H (the content of which are incorporated herein by reference in their entirety). Access to the foregoing information in the Protein Data Bank can be found at www.rcsb.org.

The R-groups of known isolated polyketide synthases can be readily determined by consulting sequence databases well known in the art, such as, for example, Genbank. Additional R-groups found inside and/or outside of the active site may or may not be the same. R-groups may be a natural R-group, unnatural R-group, hydrophobic R-group, hydrophilic R-group, positively charged R-group, negatively charged R-group, and the like. The term "mutant" only refers to the configuration of R-groups within the active site; therefore, mutations outside of the residues found in the active site are not considered to be mutants in light of the present invention.

"Nonmutated synthase" includes a synthase wherein no R-group(s) are changed relative to the active site of CHS (see, for example, PDB Accession No. 1BI5). A nonmutated synthase according to the present invention may or may not have amino acid residues outside of the active site that are the same as those taught for native CHS. In addition, a nonmutated synthase is a synthase having an active site comprising α-carbons having the coordinates as given in Table 1 and having the arrangements of R-groups associated with α-carbons as given in Table 1.

TABLE 1

Structural Cartesian coordinates of α-carbons found in the active site of a polyketide synthase of the present invention.

Active Site –Carbon Number	X Position	Y Position	Z Position	Amino Acid
1	25.378	49.320	57.979	Thr 132
2	26.089	45.704	56.981	Ser 133
3	35.423	42.296	66.622	Met 137*
4	25.212	49.977	62.196	Gln 161
5	22.745	44.120	51.193	Thr 194
6	19.022	42.892	54.600	Thr 197
7	13.850	48.144	50.791	Gly 211
8	22.118	48.048	46.357	Gly 216
9	13.001	54.666	59.688	Ile 254
10	16.434	48.819	61.334	Gly 256
11	18.715	43.328	59.526	Leu 263
12	13.943	47.516	57.567	Phe 265
13	9.252	52.715	57.456	Leu 267
14	23.141	53.552	52.148	Ser 338

* Met 137 from the second monomer

"Non-native" or "non-native synthase" refers to synthase proteins that are not found in nature, whether isolated or not. A non-native synthase may, for example, be a mutated synthase (see, for example, PDB Accession No. 1D6F, 1D6I and 1D6H).

"Native" or "native synthase" refers to synthase proteins that are produced in nature, *e.g.*, are not mutants (see, for example, PDB Accession No. 1BI5).

"Isolated" refers to a protein or nucleic acid that has been identified and separated from its natural environment. Contaminant components of its natural environment may include enzymes, hormones, and other proteinaceous or non-proteinaceous solutes. In one embodiment, the isolated molecule, in the case of a protein, will be purified to a degree sufficient to obtain at least 15 residues of

N-terminal or internal amino acid sequence or to homogeneity by SDS-PAGE under reducing or non-reducing conditions using Coomassie blue or silver stain. In the case of a nucleic acid the isolated molecule will preferably be purified to a degree sufficient to obtain a nucleic acid sequence using standard sequencing methods.

"Degenerate variations thereof" refers to changing a gene sequence using the degenerate nature of the genetic code to encode proteins having the same amino acid sequence yet having a different gene sequence. For example, polyketide synthases of the present invention are based on amino acid sequences. Degenerate gene variations thereof can be made encoding the same protein due to the plasticity of the genetic code, as described herein.

"Expression" refers to transcription of a gene or nucleic acid sequence, stable accumulation of nucleic acid, and the translation of that nucleic acid to a polypeptide sequence. Expression of genes also involves transcription of the gene to make RNA, processing of RNA into mRNA in eukaryotic systems, and translation of mRNA into proteins. It is not necessary for the genes to integrate into the genome of a cell in order to achieve expression. This definition in no way limits expression to a particular system or to being confined to cells or a particular cell type and is meant to include cellular, transient, *in vitro, in vivo*, and viral expression systems in both prokaryotic, eukaryotic cells, and the like.

"Foreign" or "heterologous" genes refers to a gene encoding a protein whose exact amino acid sequence is not normally found in the host cell.

"Promoter" and "promoter regulatory element", and the like, refers to a nucleotide sequence element within a nucleic acid fragment or gene that controls the expression of that gene. These can also include expression control sequences. Promoter regulatory elements, and the like, from a variety of sources can be used efficiently to promote gene expression. Promoter regulatory elements are meant to include constitutive, tissue-specific, developmental-specific, inducible, subgenomic promoters, and the like. Promoter regulatory elements may also include certain enhancer elements or silencing elements that improve or regulate transcriptional

efficiency. Promoter regulatory elements are recognized by RNA polymerases, promote the binding thereof, and facilitate RNA transcription.

A polypeptide is a chain of amino acids, regardless of length or post-translational modification (*e.g.*, glycosylation or phosphorylation). A polypeptide or protein refers to a polymer in which the monomers are amino acid residues, which are joined together through amide bonds. When the amino acids are alpha-amino acids, either the L-optical isomer or the D-optical isomer can be used, the L-isomers being typical. A synthase polypeptide of the invention is intended to encompass an amino acid sequence as set forth in SEQ ID NO:1 (see, Table 2) or SEQ ID NO:1 having one or more of the following mutations: C164A, H303Q, and N336A, mutants, variants and conservative substitutions thereof comprising L- or D- amino acids and include modified sequences such as glycoproteins.

TABLE 2 (SEQ ID NO:1)

```
MVSVSEIRKA QRAEGPATIL AIGTANPANC VEQSTYPDFY FKITNSEHKT ELKEKFQRMC
DKSMIKRRYM YLTEEILKEN PNVCEYMAPS LDARQDMVVV EVPRLGKEAA VKAIKEWGQP
KSKITHLIVC TTSGVDMPGA DYQLTKLLGL RPYVKRYMMY QQGXFAGGTV LRLAKDLAEN
NKGARVLVVC SEVTAVTFRG PSDTHLDSLV GQALFGDGAA ALIVGSDPVP EIEKPIFEMV
WTAQTIAPDS EGAIDGHLRE AGLTFHLLKD VPGIVSKNIT KALVEAFEPL GISDYNSIFW
IAHPGGPAIL DQVEQKLALK PEKMNATREV LSEYGNMSSA CVLFILDEMR KKSTQNGLKT
TGEGLEWGVL FGFGPGLTIE TVVLRSVAI
```

Accordingly, the polypeptides of the invention are intended to cover naturally occurring proteins, as well as those which are recombinantly or synthetically synthesized. Polypeptide or protein fragments are also encompassed by the invention. Fragments can have the same or substantially the same amino acid sequence as the naturally occurring protein. A polypeptide or peptide having substantially the same sequence means that an amino acid sequence is largely, but not entirely, the same, but retains a functional activity of the sequence to which it is related. In general polypeptides of the invention include peptides, or full-length protein, that contains substitutions, deletions, or insertions into the protein backbone, that would still have an

approximately 70%-90% homology to the original protein over the corresponding portion. A yet greater degree of departure from homology is allowed if like-amino acids, *i.e.* conservative amino acid substitutions, do not count as a change in the sequence.

A polypeptide may be substantially related but for a conservative variation, such polypeptides being encompassed by the invention. A conservative variation denotes the replacement of an amino acid residue by another, biologically similar residue. Examples of conservative variations include the substitution of one hydrophobic residue such as isoleucine, valine, leucine or methionine for another, or the substitution of one polar residue for another, such as the substitution of arginine for lysine, glutamic for aspartic acids, or glutamine for asparagine, and the like. Other illustrative examples of conservative substitutions include the changes of: alanine to serine; arginine to lysine; asparagine to glutamine or histidine; aspartate to glutamate; cysteine to serine; glutamine to asparagine; glutamate to aspartate; glycine to proline; histidine to asparagine or glutamine; isoleucine to leucine or valine; leucine to valine or isoleucine; lysine to arginine, glutamine, or glutamate; methionine to leucine or isoleucine; phenylalanine to tyrosine, leucine or methionine; serine to threonine; threonine to serine; tryptophan to tyrosine; tyrosine to tryptophan or phenylalanine; valine to isoleucine or leucine, and the like. The term "conservative variation" also includes the use of a substituted amino acid in place of an unsubstituted parent amino acid provided that antibodies raised to the substituted polypeptide also immunoreact with the unsubstituted polypeptide.

Modifications and substitutions are not limited to replacement of amino acids. For a variety of purposes, such as increased stability, solubility, or configuration concerns, one skilled in the art will recognize the need to introduce, (by deletion, replacement, or addition) other modifications. Examples of such other modifications include incorporation of rare amino acids, dextra-amino acids, glycosylation sites, cytosine for specific disulfide bridge formation. The modified peptides can be chemically synthesized, or the isolated gene can be site-directed mutagenized, or a

synthetic gene can be synthesized and expressed in bacteria, yeast, baculovirus, tissue culture and so on.

Chalcone synthase polypeptides of the invention include synthase polypeptides from plants, prokaryotes, eukaryotes, including, for example, invertebrates, mammals and humans and include sequences as set forth in SEQ ID NO:1, as well as sequences that have at least 70% homology to the sequence of SEQ ID NO:1, fragments, variants, or conservative substitutions of any of the foregoing sequences.

The term "variant" refers to polypeptides modified at one or more amino acid residues yet still retain the biological activity of a synthase polypeptide. Variants can be produced by any number of means known in the art, including, for example, methods such as, for example, error-prone PCR, shuffling, oligonucleotide-directed mutagenesis, assembly PCR, sexual PCR mutagenesis, and the like, as well as any combination thereof.

By "substantially identical" is meant a polypeptide or nucleic acid exhibiting at least 50%, preferably 85%, more preferably 90%, and most preferably 95% homology to a reference amino acid or nucleic acid sequence.

Homology or identity is often measured using sequence analysis software (*e.g.*, Sequence Analysis Software Package of the Genetics Computer Group, University of Wisconsin Biotechnology Center, 1710 University Avenue, Madison, WI 53705). Such software matches similar sequences by assigning degrees of homology to various deletions, substitutions and other modifications. The terms "homology" and "identity" in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same when compared and aligned for maximum correspondence over a comparison window or designated region as measured using any number of sequence comparison algorithms or by manual alignment and visual inspection.

For sequence comparison, typically one sequence acts as a reference sequence, to which test sequences are compared. When using a sequence comparison algorithm, test and reference sequences are entered into a computer, subsequence coordinates are designated, if necessary, and sequence algorithm program parameters are designated. Default program parameters can be used, or alternative parameters can be designated. The sequence comparison algorithm then calculates the percent sequence identities for the test sequences relative to the reference sequence, based on the program parameters.

A "comparison window", as used herein, includes reference to a segment of any one of the number of contiguous positions selected from the group consisting of from 20 to 600, usually about 50 to about 200, more usually about 100 to about 150 in which a sequence may be compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned. Methods of alignment of sequence for comparison are well-known in the art. Optimal alignment of sequences for comparison can be conducted, *e.g.*, by the local homology algorithm of Smith & Waterman, Adv. Appl. Math. 2:482, 1981, by the homology alignment algorithm of Needleman & Wunsch, J. Mol. Biol 48:443, 1970, by the search for similarity method of Person & Lipman, Proc. Nat'l. Acad. Sci. USA 85:2444, 1988, by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by manual alignment and visual inspection. Other algorithms for determining homology or identity include, for example, in addition to a BLAST program (Basic Local Alignment Search Tool at the National Center for Biological Information), ALIGN, AMAS (Analysis of Multiply Aligned Sequences), AMPS (Protein Multiple Sequence Alignment), ASSET (Aligned Segment Statistical Evaluation Tool), BANDS, BESTSCOR, BIOSCAN (Biological Sequence Comparative Analysis Node), BLIMPS (BLocks IMProved Searcher), FASTA, Intervals & Points, BMB, CLUSTAL V, CLUSTAL W, CONSENSUS, LCONSENSUS, WCONSENSUS, Smith-Waterman algorithm, DARWIN, Las Vegas algorithm, FNAT (Forced Nucleotide Alignment Tool), Framealign, Framesearch, DYNAMIC, FILTER, FSAP (Fristensky Sequence Analysis Package), GAP (Global

Alignment Program), GENAL, GIBBS, GenQuest, ISSC (Sensitive Sequence Comparison), LALIGN (Local Sequence Alignment), LCP (Local Content Program), MACAW (Multiple Alignment Construction & Analysis Workbench), MAP (Multiple Alignment Program), MBLKP, MBLKN, PIMA (Pattern-Induced Multi-sequence Alignment), SAGA (Sequence Alignment by Genetic Algorithm) and WHAT-IF. Such alignment programs can also be used to screen genome databases to identify polynucleotide sequences having substantially identical sequences. A number of genome databases are available, for example, a substantial portion of the human genome is available as part of the Human Genome Sequencing Project (J. Roach, http://weber.u.Washington.edu/~roach/human_ genome_ progress 2.html) (Gibbs, 1995). At least twenty-one other genomes have already been sequenced, including, for example, *M. genitalium* (Fraser *et al.*, 1995), M. jannaschii (Bult *et al.*, 1996), *H. influenzae* (Fleischmann *et al.*, 1995), *E. coli* (Blattner *et al.*, 1997), and yeast (*S. cerevisiae*) (Mewes *et al.*, 1997), and *D. melanogaster* (Adams *et al.*, 2000). Significant progress has also been made in sequencing the genomes of model organism, such as mouse, *C. elegans*, and *Arabadopsis sp*. Several databases containing genomic information annotated with some functional information are maintained by different organization, and are accessible via the internet, for example, http://wwwtigr.org/tdb; http://www.genetics.wisc.edu; http://genome-www.stanford.edu/~ball; http://hiv-web.lanl.gov; http://www.ncbi.nlm.nih.gov; http://www.ebi.ac.uk; http://Pasteur.fr/other/biology; and http:// www.genome.wi.mit.edu.

One example of a useful algorithm is BLAST and BLAST 2.0 algorithms, which are described in Altschul *et al.*, Nuc. Acids Res. 25:3389-3402, 1977, and Altschul *et al.*, J. Mol. Biol. 215:403-410, 1990, respectively. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (http://www.ncbi.nlm.nih. gov). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul *et al.*, *supra*). These initial neighborhood

word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always >0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a wordlength (W) of 11, an expectation (E) of 10, M=5, N=-4 and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength of 3, and expectations (E) of 10, and the BLOSUM62 scoring matrix (see Henikoff & Henikoff, Proc. Natl. Acad. Sci. USA 89:10915, 1989) alignments (B) of 50, expectation (E) of 10, M=5, N= -4, and a comparison of both strands.

The BLAST algorithm also performs a statistical analysis of the similarity between two sequences (see, *e.g.*, Karlin & Altschul, Proc. Natl. Acad. Sci. USA 90:5873, 1993). One measure of similarity provided by BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a references sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.2, more preferably less than about 0.01, and most preferably less than about 0.001.

In one embodiment, protein and nucleic acid sequence homologies are evaluated using the Basic Local Alignment Search Tool ("BLAST") In particular, five specific BLAST programs are used to perform the following task:

(1) BLASTP and BLAST3 compare an amino acid query sequence against a protein sequence database;

(2) BLASTN compares a nucleotide query sequence against a nucleotide sequence database;

(3) BLASTX compares the six-frame conceptual translation products of a query nucleotide sequence (both strands) against a protein sequence database;

(4) TBLASTN compares a query protein sequence against a nucleotide sequence database translated in all six reading frames (both strands); and

(5) TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

The BLAST programs identify homologous sequences by identifying similar segments, which are referred to herein as "high-scoring segment pairs," between a query amino or nucleic acid sequence and a test sequence which is preferably obtained from a protein or nucleic acid sequence database. High-scoring segment pairs are preferably identified (*i.e.*, aligned) by means of a scoring matrix, many of which are known in the art. Preferably, the scoring matrix used is the BLOSUM62 matrix (Gonnet *et al.*, Science 256:1443-1445, 1992; Henikoff and Henikoff, Proteins 17:49-61, 1993). Less preferably, the PAM or PAM250 matrices may also be used (see, *e.g.*, Schwartz and Dayhoff, eds., 1978, *Matrices for Detecting Distance Relationships: Atlas of Protein Sequence and Structure*, Washington: National Biomedical Research Foundation). BLAST programs are accessible through the U.S. National Library of Medicine, *e.g.*, at www.ncbi.nlm.nih.gov.

The parameters used with the above algorithms may be adapted depending on the sequence length and degree of homology studied. In some embodiments, the

parameters may be the default parameters used by the algorithms in the absence of instructions from the user.

By a "substantially pure polypeptide" is meant a synthase polypeptide (e.g., a chalcone synthase) which has been separated from components which naturally accompany it. Typically, the polypeptide is substantially pure when it is at least 60%, by weight, free from the proteins and naturally-occurring organic molecules with which it is naturally associated. Preferably, the preparation is at least 75%, more preferably at least 90%, and most preferably at least 99%, by weight, synthase polypeptide. A substantially pure synthase polypeptide may be obtained, for example, by extraction from a natural source; by expression of a recombinant nucleic acid encoding an synthase polypeptide; or by chemically synthesizing the protein. Purity can be measured by any appropriate method (*e.g.*, column chromatography, polyacrylamide gel electrophoresis, or by HPLC analysis).

One aspect of the invention resides in obtaining crystals of the synthase polypeptide chalcone synthase of sufficient quality to determine the three dimensional (tertiary) structure of the protein by X-ray diffraction methods. The knowledge obtained concerning the three-dimensional structure of chalcone synthase can be used in the determination of the three dimensional structure of other synthase polypeptides in the polyketide synthesis pathway. The structural coordinates of chalcone synthase can be used to develop new polyketide synthesis enzymes or synthase inhibitors using various computer models. Based on the structural coordinates of the chalcone synthase polypeptide (*e.g.*, the three dimensional protein structure), as described herein, novel polyketide synthases can be engineered. In addition, small molecules which mimic or are capable of interacting with a functional domain of a synthase molecule, can be designed and synthesized to modulate chalcone synthase, pyrone synthase, and other polyketide synthase biological functions as well as the biological functions of other polyketide synthases. Accordingly, in one embodiment, the invention provides a method of "rational" enzyme or drug design. Another approach

to "rational" enzyme or drug design is based on a lead compound that is discovered using high throughput screens; the lead compound is further modified based on a crystal structure of the binding regions of the molecule in question. Accordingly, another aspect of the invention is to provide related protein sequences or material which is a starting material in the rational design of new synthases or drugs which lead to the synthesis of new polyketides or modify the polyketide synthesis pathway.

"Active Site" refers to a site in a synthase defined by amino acid residues that interact with substrate and facilitate a biosynthetic reaction that allows one or more products to be produced. An active site is comprised of α-carbon atoms that are indirectly linked via peptide bonds and have the structural coordinates disclosed in Table 1 ± 2.3 angstroms. Other active site amino acids for chalcone synthase include C164, H303, and N336. The position in three-dimensional space of an α-carbon at the active site of a synthase and of R-groups associated therewith can be determined using techniques such as three-dimensional modeling, X-ray crystallography, and/or techniques associated therewith.

"Altered substrate specificity" includes a change in the ability of a mutant synthase to produce a polyketide product as compared to a non-mutated synthase. Altered substrate specificity may include the ability of a synthase to exhibit different enzymatic parameters relative to a non-mutated synthase (K_m, V_{max}. etc), use different substrates, and/or produce products that are different from those of known non-native synthases.

"Structure coordinates" refers to Cartesian coordinates (x, y, and z positions) derived from mathematical equations involving Fourier synthesis as determined from patterns obtained via diffraction of a monochromatic beam of X-rays by the atoms (scattering centers) of a polyketide synthase molecule in crystal form. Diffraction data are used to calculate electron density maps of repeating protein units in the crystal (unit cell). Electron density maps are used to establish the positions of individual atoms within a crystal's unit cell. The term "crystal structure coordinates" refers to mathematical coordinates derived from mathematical equations related to the

patterns obtained on diffraction of a monochromatic beam of X-rays by the atoms (scattering centers) of a synthase polypeptide (*e.g.*, a chalcone synthase protein molecule) in crystal form. The diffraction data are used to calculate an electron density map of the repeating unit of the crystal. The electron density maps are used to establish the positions of the individual atoms within the unit cell of the crystal. The crystal structure coordinates of a synthase can be obtained from a chalcone synthase protein crystal having space group $P3_221$ ($a = b = 97.54$ Å, $c = 65.52$ with a single monomer per asymmetric unit). The coordinates of the cynthase polypeptide can also be obtained by means of computational analysis.

The term "selenomethionine substitution" refers to the method of producing a chemically modified form of the crystal of a synthase (e.g., a chalcone synthase). The synthase protein is expressed by bacteria in media that is depleted in methionine and supplement with selenomethionine. Selenium is thereby incorporated into the crystal in place of methionine sulfurs. The location(s) of selenium are determined by X-ray diffraction analysis of the crystal. This information is used to generate the phase information used to construct a three-dimensional structure of the protein.

"Heavy atom derivatization" refers to a method of producing a chemically modified form of a synthase crystal. In practice, a crystal is soaked in a solution containing heavy atom salts or organometallic compounds, *e.g.*, lead chloride, gold thiomalate, thimerosal, uranyl acetate, and the like, which can diffuse through the crystal and bind to the protein's surface. Locations of the bound heavy atoms can be determined by X-ray diffraction analysis of the soaked crystal. This information is then used to construct phase information which can then be used to construct three-dimensional structures of the enzyme as described in Blundel, T. L., and Johnson, N. L., Protein Crystallography, Academic Press (1976), which is incorporated herein by reference.

"Unit cell" refers to a basic parallelepiped shaped block. Regular assembly of such blocks may construct the entire volume of a crystal. Each unit cell comprises a complete representation of the unit pattern, the repetition of which builds up the

crystal.

"Mutagenesis" refers to the changing of one R-group for another as defined herein. This can be most easily performed by changing the coding sequence of the nucleic acid encoding the amino acid residue. In the context of the present invention, mutagenesis does not change the carbon coordinates beyond the limits defined herein.

"Space Group" refers to the arrangement of symmetry elements within a crystal.

"Molecular replacement" refers to generating a preliminary model of a polyketide synthase whose structural coordinates are unknown, by orienting and positioning a molecule whose structural coordinates are known within the unit cell of the unknown crystal so as best to account for the observed diffraction pattern of the unknown crystal. Phases can then be calculated from this model and combined with the observed amplitudes to give an approximate Fourier synthesis of the structure whose coordinates are unknown. This in turn can be subject to any of the several forms of refinement to provide a final, accurate structure of the unknown crystal (Lattman, E., 1985, in Methods in Enzymology, 11 5.55-77; Rossmann, MG., ed., "The Molecular Replacement Method" 1972, Int, Sci. Rev. Ser., No. 13, Gordon & Breach, New York). Using structure coordinates of the polyketide synthase provided in Figure 1 molecular replacement may be used to determine the structural coordinates of a crystalline mutant, homologue, or a different crystal form of polyketide synthase.

A "synthase" or a "polyketide synthase" includes any one of a family of enzymes that catalyze the formation of polyketide compounds. Polyketide synthases are generally homodimers, with each monomer being enzymatically acitve.

"Substrate" refers to the Coenzyme-A (CoA) thioesters that are acted on by the polyketide synthases and mutants thereof disclosed herein, such as malonyl-CoA, coumaroyl-CoA, hexamoyl-CoA, and the like.

The present invention relates to crystallized polyketide synthases and mutants thereof from which the position of specific α-carbon atoms and R-groups associated therewith comprising the active site can be determined in three-dimensional space.

The invention also relates to structural coordinates of said polyketide synthases, use of said structural coordinates to develop structural information related to polyketide synthase homologues, mutants, and the like, and to crystal forms of such synthases. Furthermore, the invention, as disclosed herein, provides a method whereby said α-carbon structural coordinates specifically determined for atoms comprising the active site of said synthase, as shown in Table 1 and including C164, H303, and N336, can be used to develop synthases wherein R-groups associated with active site α-carbon atoms are different from the R-groups found in native CHS, *e.g.*, are mutant synthases. In addition, the present invention provides for production of mutant polyketide synthases based on the structural information provided herein and for use of said mutant synthases to make a variety of polyketide compounds using a variety of substrates.

The present invention further provides, for the first time, crystals of several polyketide synthases, as exemplified by chalcone synthase (CHS; PDB Accession No. 1B15), stilbene synthase (STS), and pyrone synthase (PS); see Table 3 for coordinates of PS ("molecule" denoted in the table refers to the particular monomer of the PS dimer). Also provided are coordinates for crystals which are grown in the presence and absence of substrate and substrate analogues, thus allowing definition of the structural or atomic coordinates associated therewith. Said structural coordinates allow determination of the carbon atoms comprising the active site, R-groups associated therewith, and the interaction of said α-carbons and said R-groups with each other. For example, Table 4 identifies various substrates and substrate analogues that were grown with chalcone synthase as well as their PDB accession numbers, all of which are incorporated herein by reference in their entirety.

TABLE 3

Atom	Atom Type	Res.	#	X	Y	Z	OCC	B	Molec
1	N	GLY	20	32.834	42.457	65.617	1.00	27.09	A
2	CA	GLY	20	33.866	41.428	65.356	1.00	23.93	A
3	C	GLY	20	33.175	40.130	64.906	1.00	21.83	A
4	O	GLY	20	31.967	40.073	64.809	1.00	20.10	A
5	N	LEU	21	34.001	39.120	64.701	1.00	19.92	A
6	CA	LEU	21	33.519	37.812	64.301	1.00	21.97	A

TABLE 4

Complex	PDB Accession No.
CHS-coA complex	1BQ6
CHS-malonyl-CoA complex	1CM1
CHS-hexanoyl-CoA comlex	1CHW
CHS-naringenin complex	1CGK
CHS-resveratrol complex	1CGZ

The crystals of the present invention belong to the tetragonal space group. The unit cell dimensions vary by a few angstroms between crystals but on average, chalcone synthase native crystals belong to space group $P3_221$ with unit cell dimensions of $a = b = 97.54$ Å; $c = 65.52$ Å, $\alpha = \beta = 90°$, $\gamma = 120°$ with a single monomer per asymmetric unit. Stilbene synthase crystals belong to space group C222 with unit cell dimensions of $a = 74.94$ Å , $b = 86.63$ Å , $c = 364.18$ Å , $\alpha = \beta = \gamma = 90°$. Pyrone synthase crystals belong to space group P3121 with unit cell dimensions of $a = 82.15$ Å, $b = 241.33$ Å, $\alpha = \beta = 90°$, $\gamma = 120°$ with one PS dimer per asymmetric unit.

Crystal structures are preferably obtained at a resolution of about 1.56 angstroms to about 3 angstroms for a polyketide synthase in the presence and in the absence of bound substrate or substrate analog. Coordinates for a polyketide synthase in the absence of a substrate bound in the active site have been deposited at the Brookhaven National Laboratory Protein Data Bank, accession number 1CGK. Those skilled in the art understand that a set of structure coordinates determined by X-ray crystallography is not without standard error. Therefore, for the purpose of this invention, any set of structure coordinates wherein the active site α-carbons of a polyketide synthase, synthase homologue, or mutants thereof, have a root mean square deviation less than ± 2.3 angstroms when superimposed using the structural coordinates listed in Table 1 and PDB Accession No. 1BI5, shall be considered

identical.

A schematic representation of the three-dimensional shape of a CHS homodimer is shown in Figure 2a, which was prepared by MOLSCRIPT (Kraulis, J. Appl. Crystallogr. 24:946-950, 1991). CHS functions as a homodimer of two 42 kDa polypeptides. The structure of CHS reveals that the enzyme forms a symmetric dimer with each monomer related by a 2-fold crystallographic axis. The dimer interface buries approximately 1580 angstroms with interactions occurring along a fairly flat surface. Two distinct structural features delineate the ends of this interface. First, the N-terminal helix of monomer A entwines with the corresponding helix of monomer B. Second, a tight loop containing a cis-peptide bond between Met_{137} and Pro_{138} exposes the methionine sidechain as a knob on the monomer surface. Across the interface, Met_{137} protrudes into a hole found in the surface of the adjoining monomer to form part of the cyclization pocket (discussed below).

The CHS homodimer contains two functionally independent active sites (Tropf, et al, J. Biol. Chem. 270:7922-7928, 1995). Consistent with this information, bound CoA thioesters and product analogs occupy both active sites of the homodimer in the CHS complex structures. These structures identify the location of the active site at the cleft between the upper and lower domains of each monomer. Each active site consists almost entirely of residues from a single monomer, with Met_{137} from the adjoining monomer being the only exception. A detailed description of the active site structure is presented in the Examples section, below.

An isolated, polyketide synthase of the invention comprises at least fourteen active site α-carbons having the structural coordinates of Table 1 ±2.3 angstroms. The active site α-carbons of Table 1 generally are not all contiguous, i.e., are not adjacent to one another in the primary amino acid sequence of a polyketide synthase due to intervening amino acid residues between various active site α-carbons. Nevertheless, it should be appreciated that certain active site α-carbons can be adjacent to one another in some instances. Active site α-carbons are numbered in Table 1 for convenience only and may be situated in any suitable order in the primary amino acid

sequence that achieves the structural coordinates given in Table 1.

An appropriate combination of R-groups, linked to active site α-carbons, can facilitate the formation of one or more desired reaction products. The combination of R-groups selected for use in a synthase can be any combination other than the ordered arrangements of R-groups found in known native isolated polyketide synthases. Typically, R-groups found on active site α-carbons are those found in naturally occurring amino acids. In some embodiments, however, R-groups other than those found in naturally occurring amino acids can be used.

The present invention permits the use of molecular design techniques to design, select, and synthesize genes encoding mutant polyketide synthases that produce different and/or novel polyketide compounds using substrates. Mutant proteins of the present invention and nucleic acids encoding the same can be designed by genetic manipulation based on structural information about polyketide synthases. For example, one or more R-groups associated with the active site α-carbon atoms of CHS can be changed by altering the nucleotide sequence of the corresponding CHS gene, thus making one or more mutant polyketide synthases. Such genetic manipulations can be guided by structural information concerning the R-groups found in the active site α-carbons when substrate is bound to the protein upon crystallization.

Mutant proteins of the present invention may be prepared in a number of ways available to the skilled artisan. For example, the gene encoding wild-type CHS may be mutated at those sites identified herein as corresponding to amino acid residues identified in the active site by means currently available to the artisan skilled in molecular biology techniques. Said techniques include oligonucleotide-directed mutagenesis, deletion, chemical mutagenesis, and the like. The protein encoded by the mutant gene is then produced by expressing the gene in, for example, a bacterial or plant expression system.

Alternatively, polyketide synthase mutants may be generated by site specific-replacement of a particular amino acid with an unnaturally occurring amino acid. As such, polyketide synthase mutants may be generated through replacement of an amino

acid residue or a particular cysteine or methionine residue with selenocysteine or selenomethionine. This may be achieved by growing a host organism capable of expressing either the wild-type or mutant polypeptide on a growth medium depleted of natural cysteine or methionine or both and growing on medium enriched with either selenocysteine, selenomethionine, or both. These and similar techniques are described in Sambrook et al., (Molecular Cloning, A Laboratory Manual, 2nd Ed. (1989) Cold Spring Harbor Laboratory Press).

Another suitable method of creating mutant synthases of the present invention is based on a procedure described in Noel and Tsal (1989) J. Cell. Biochem., 40:309-320. In so doing, the nucleic acids encoding said polyketide synthase can be synthetically produced using oligonucleotides having overlapping regions, said oligonucleotides being degenerate at specific bases so that mutations are induced.

According to the present invention, nucleic acid sequences encoding a mutated polyketide synthase can be produced by the methods described herein, or any alternative methods available to the skilled artisan. In designing the nucleic acid sequence of interest, it may be desirable to reengineer said gene for improved expression in a particular expression system. For example, it has been shown that many bacterially derived genes do not express well in plant systems. In some cases, plant-derived genes do not express well in bacteria. This phenomenon may be due to the non-optimal G+C content and/or A+T content of said gene relative to the expression system being used. For example, the very low G+C content of many bacterial genes results in the generation of sequences mimicking or duplicating plant gene control sequences that are highly A+T rich. The presence of A+T rich sequences within the genes introduced into plants (e.g., TATA box regions normally found in promoters) may result in aberrant transcription of the gene(s). In addition, the presence of other regulatory sequences residing in the transcribed mRNA (e.g. polyadenylation signal sequences (AAUAAA) or sequences complementary to small nuclear RNAs involved in pre-mRNA splicing) may lead to RNA instability. Therefore, one goal in the design of genes is to generate nucleic acid sequences that have a G+C content that affords mRNA stability and translation accuracy for a

particular expression system.

Due to the plasticity afforded by the redundancy of the genetic code (i.e., some amino acids are specified by more than one codon), evolution of the genomes of different organisms or classes of organisms has resulted in differential usage of redundant codons. This "codon bias" is reflected in the mean base composition of protein coding regions. For example, organisms with relatively low G+C contents utilize codons having A or T in the third position of redundant codons, whereas those having higher G+C contents utilize codons having G or C in the third position. Therefore, in reengineering genes for expression, one may wish to determine the codon bias of the organism in which the gene is to be expressed. Looking at the usage of the codons as determined for genes of a particular organism deposited in GenBank can provide this information. After determining the bias thereof, the new gene sequence can be analyzed for restriction enzyme sites as well as other sites that could affect transcription such as exon:intron junctions, polyA addition signals, or RNA polymerase termination signals.

Genes encoding polyketide synthases can be placed in an appropriate vector, depending on the artisan's interest, and can be expressed using a suitable expression system. An expression vector, as is well known in the art, typically includes elements that permit replication of said vector within the host cell and may contain one or more phenotypic markers for selection of cells containing said gene. The expression vector will typically contain sequences that control expression such as promoter sequences, ribosome binding sites, and translational initiation and termination sequences. Expression vectors may also contain elements such as subgenomic promoters, a repressor gene or various activator genes. The artisan may also choose to include nucleic acid sequences that result in secretion of the gene product, movement of said product to a particular organelle such as a plant plastid (see U.S. Patent Nos. 4,762,785; 5,451,513 and 5,545,817, which are incorporated herein by reference) or other sequences that increase the ease of peptide purification, such as an affinity tag.

A wide variety of expression control sequences are useful in expressing the

mutated polyketide synthases when operably linked thereto. Such expression control sequences include, for example, the early and late promoters of SV40 for animal cells, the lac system, the trp system, major operator and promoter systems of phage S, and the control regions of coat proteins, particularly those from RNA viruses in plants. In *E. coli*, a useful transcriptional control sequence is the T7 RNA polymerase binding promoter, which can be incorporated into a pET vector as described by Studier et al., (1990) Methods Enzymology, 185:60-89, which is incorporated herein by reference.

For expression, a desired gene should be operably linked to the expression control sequence and maintain the appropriate reading frame to permit production of the desired polyketide synthase. Any of a wide variety of well-known expression vectors are of use to the present invention. These include, for example, vectors comprising segments of chromosomal, non-chromosomal and synthetic DNA sequences such as those derived from SV40, bacterial plasmids including those from *E. coli* such as col E1, pCR1, pBR322 and derivatives thereof, pMB9), wider host range plasmids such as RP4, phage DNA such as phage S, NM989, M13, and other such systems as described by Sambrook et al., (Molecular Cloning, A Laboratory Manual, 2nd Ed. (1989) Cold Spring Harbor Laboratory Press), which is incorporated herein by reference.

A wide variety of host cells are available for expressing synthase mutants of the present invention. Such host cells include, for example, bacteria such as *E. coli*, *Bacillus* and *Streptomyces*, fungi, yeast, animal cells, plant cells, insect cells, and the like. Preferred embodiments of the present invention include chalcone synthase mutants that are expressed in *E. coli* or in plant cells. Said plant cells can either be in suspension culture or a transgenic plant as further described herein.

As stated previously, genes encoding synthases of the present invention can be expressed in transgenic plant cells. In order to produce transgenic plants, vectors containing the nucleic acid construct encoding polyketide synthases and mutants thereof are inserted into the plant genome. Preferably, these recombinant vectors are

capable of stable integration into the plant genome. One variable in making a transgenic plant is the choice of a selectable marker. A selectable marker is used to identify transformed cells against a high background of untransformed cells. The preference for a particular marker is at the discretion of the artisan, but any of the selectable markers may be used along with any other gene not listed herein that could function as a selectable marker. Such selectable markers include aminoglycoside phosphotransferase gene of transposon Tn5 (Aph 11) (which encodes resistance to the antibiotics kanamycin), neomycin, G418, as well as those genes which code for resistance or tolerance to glyphosate, hygromycin, methotrexate, phosphinothricin, imidazolinones, sulfonylureas, triazolophyrimidine herbicides, such as chlorosulfuron, bromoxynil, dalapon, and the like. In addition to a selectable marker, it may be desirable to use a reporter gene. In some instances a reporter gene may be used with a selectable marker. Reporter genes allow the detection of transformed cells and may be used at the discretion of the artisan. A list of these reporter genes is provided in K. Wolsing et al., 1988, Ann. Rev. Genetics, 22:421.

Said genes are expressed either by promoters expressing in all tissues at all times (constitutive promoters), by promoters expressing in specific tissues (tissue-specific promoters), promoters expressing at specific stages of development (developmental promoters), and/or promoter expression in response to a stimulus or stimuli (inducible promoters). The choice of these is at the discretion of the artisan.

Several techniques exist for introducing foreign genes into plant cells, and for obtaining plants that stably maintain and express the introduced gene. Such techniques include acceleration of genetic material coated on a substrate directly into cells (U.S. Patents 4,945,050 to Comell): Plant cells may also be transformed using *Agrobacterium* technology (see, for example, U.S. Patents 5,177,010 to University of Toledo, 5,104,310 to Texas A&M, U. S. Patents 5,149,645, 5,469,976, 5,464,763, 4,940,838, and 4,693,976 to Schilperoot, European Patent Applications 116718, 290799, 320500 to Max Planck, European Patent Applications 604662,627752 and U.S. Patent 5,591,616 to Japan Tobacco, European Patent Applications 0267159, 0292435 and U.S. Patent 5,231,01.q to Ciba-Geigy, U.S. Patents 5,463,174 and

4,762,785 to Calgene, and U.S. Patents 5,004,863 and 5,159,135 to Agracetus). Other transformation technologies include whiskers technology (see U. S. Patents 5,302,523 and 5,464,765 to Zeneca). Electroporation technology has also been used to transform plants (see WO 87106614 to Boyce Thompson Institute, 5,472,869 and 5,384,253 to Dakalb, and WO 92/09696 and WO 93/21335 to Plant Genetic Systems, all which are incorporated by reference). Viral vector expression systems can also be used such as those described in U.S, Patent 5,316,931, 5,589,367, 5,811,653, and 5,866,785 to BioSource, which are incorporated herein by reference.

In addition to numerous technologies for transforming plants, the type of tissue that is contacted with the genes of interest may vary as well. Suitable tissue includes, for example, embryonic tissue, callus tissue, hypocotyl, meristem, and the like. Almost all plant tissues may be transformed during de-differentiation using the appropriate techniques described herein.

Regardless of the transformation system used, a gene encoding a mutant polyketide synthase is preferably incorporated into a gene transfer vector adapted to express said gene in a plant cell by including in the vector an expression control sequence (plant promoter regulatory element). In addition to plant promoter regulatory elements, promoter regulatory elements from a variety of sources can be used efficiently in plant cells to express foreign genes. For example, promoter regulatory elements of bacterial origin, such as the octopine synthase promoter, the nopaline synthase promoter, the mannopine synthase promoter, and the like, may be used. Promoters of viral origin, such as the cauliflower mosaic virus (35S and 198) are also desirable. Plant promoter regulatory elements also include ribulose-1,6-bisphosphate carboxylase small subunit promoter, beta-conglycinin promoter, phaseolin promoter, ADH promoter, heat-shock promoters, tissue specific promoters, and the like. Numerous promoters are available to skilled artisans for use at their discretion.

It should be understood that not all expression vectors and expression systems function in the same way to express the mutated gene sequences of the present

invention. Neither do all host cells function equally well with the same expression system. However, one skilled in the art may make a selection among these vectors, expression control sequences, and host without undue experimentation and without departing from the scope of this invention.

Once a synthase of the present invention is expressed, the protein obtained therefrom can be purified so that structural analysis, modeling, and/or biochemical analysis can be performed, as exemplified herein. The nature of the protein obtained can be dependent on the expression system used. For example, genes, when expressed in mammalian or other eukaryotic cells, may contain latent signal sequences that may result in glycosylation, phosphorylation, or other post-translational modifications, which may or may not alter function. Therefore, a preferred embodiment of the present invention is the expression of mutant synthase genes in *E. coli* calls. Once said proteins are expressed, they can be easily purified using techniques common to the person having ordinary skill in the art of protein biochemistry, such as, for example, techniques described in Colligan at al., (1997) Current Protocols in Protein Science, Chanda, V. B., Ed., John Wiley & Sons, Inc., which is incorporated herein by reference. Such techniques often include the use of cation-exchange or anion-exchange chromatography, gel filtration-size exclusion chromatography, and the like. Another technique that may be commonly used is affinity chromatography. Affinity chromatography can include the use of antibodies, substrate analogs, or histidine residues (His-tag technology).

Once purified, mutants of the present invention may be characterized by any of several different properties. For example, such mutants may have altered active site surface charges of one or more charge units. In addition, said mutants may have altered substrate specificity or product capability relative to a non-mutated polyketide synthase.

The present invention allows for the characterization of polyketide synthase mutants by crystallization followed by X-ray diffraction. Polypeptide crystallization occurs in solutions where the polypeptide concentration exceeds it solubility

maximum (i.e., the polypeptide solution is supersaturated). Such solutions may be restored to equilibrium by reducing the polypeptide concentration, preferably through precipitation of the polypeptide crystals. Often polypeptides may be induced to crystallize from supersaturated solutions by adding agents that alter the polypeptide surface charges or perturb the interaction between the polypeptide and bulk water to promote associations that lead to crystallization.

Compounds known as "precipitants" are often used to decrease the solubility of the polypeptide in a concentrated solution by forming an energetically unfavorable precipitating layer around the polypeptide molecules (Weber, Advances in Protein Chemistry, 41:1-36, 1991). In addition to precipitants, other materials are sometimes added to the polypeptide crystallization solution. These include buffers to adjust the pH of the solution and salts to reduce the solubility of the polypeptide. Various precipitants are known in the art and include the following: ethanol, 3-ethyl-2-4 pentanediol, and many of the polyglycols, such as polyethylene glycol.

Commonly used polypeptide crystallization methods include the following techniques: batch, hanging drop, seed initiation, and dialysis. In each of these methods, it is important to promote continued crystallization after nucleation by maintaining a supersaturated solution. In the batch method, polypeptide is mixed with precipitants to achieve supersaturation, the vessel is sealed, and set aside until crystals appear. In the dialysis method, polypeptide is retained in a sealed dialysis membrane that is placed into a solution containing precipitant. Equilibration across the membrane increases the polypeptide and precipitant concentrations thereby causing the polypeptide to reach supersaturation levels.

In the preferred hanging drop technique (McPherson, J. Biot Chem, 6300-6306, 1976), an initial polypeptide mixture is created by adding a precipitant to a concentrated polypeptide solution. The concentrations of the polypeptide and precipitants are such that in this initial form, the polypeptide does not crystallize. A small drop of this mixture is placed on a glass slide that is inverted and suspended

over a reservoir of a second solution. The system is then sealed. Typically, the second solution contains a higher concentration of precipitant or other dehydrating agent. The difference in the precipitant concentrations causes the protein solution to have a higher vapor pressure than the solution. Since the system containing the two solutions is sealed, an equilibrium is established, and water from the polypeptide mixture transfers to the second solution. This equilibrium increases the polypeptide and precipitant concentration in the polypeptide solution. At the critical concentration of polypeptide and precipitant, a crystal of the polypeptide will form.

Another method of crystallization introduces a nucleation site into a concentrated polypeptide solution. Generally, a concentrated polypeptide solution is prepared and a seed crystal of the polypeptide is introduced into this solution. If the concentration of the polypeptide and any precipitants are correct, the seed crystal will provide a nucleation site around which a larger crystal forms. In preferred embodiments, the crystals of the present invention are formed in hanging drops with (15% PEG 8000; 200 mM magnesium acetate or magnesium chloride, 100 mM 3-(N-morpholino)-2-hydroxypropanesulfonic acid (pH 7.0), 1 mM dithiothreitol as precipitant).

Some proteins may be recalcitrant to crystallization. However, several techniques are available to the skilled artisan. Quite often the removal of polypeptide segments at the amino or caroxy terminal end of the protein is necessary to produce crystalline protein samples. Said procedures involve either the treatment of the protein with one of several proteases including trypsin, chymotrypsin, substilisin, and the like. This treatment often results in the removal of flexible polypeptide segments that are likely to negatively affect crystallization. Alternatively, the removal of coding sequences from the protein's gene facilitates the recombinant expression of shortened proteins that can be screened for crystallization.

The crystals so produced have a wide range of uses. For example, high quality crystals are suitable for X-ray or neutron diffraction analysis to determine the three-dimensional structure of a mutant polyketide synthase and to design additional

mutants thereof. In addition, crystallization can serve as a further purification method. In some instances, a polypeptide or protein will crystallize from a heterogeneous mixture into crystals. Isolation of such crystals by filtration, centrifugation, etc., followed by redissolving the polypeptide affords a purified solution suitable for use in growing the high-quality crystals needed for diffraction studies. The high-quality crystals may also be dissolved in water and then formulated to provide an aqueous solution having other uses as desired.

Because synthases may crystallize in more than one crystal form, the structural coordinates of α-carbons of an active site determined from a synthase or portions thereof, as provided by this invention, are particularly useful to solve the structure of other crystal forms of synthases. Said structural coordinates, as provided herein, may also be used to solve the structure of synthases having α-carbons positioned within the active sites in a manner similar to the wild-type, yet having R-groups that may or may not be identical.

Furthermore, the structural coordinates disclosed herein may be used to determine the structure of the crystalline form of other proteins with significant amino acid or structural homology to any functional domain of a synthase. One method that may be employed for such purpose is molecular replacement. In this method, the unknown crystal structure, whether it is another crystal form of a synthase, a synthase having a mutated active site, or the crystal of some other protein with significant sequence and/or structural homology to a polyketide synthase may be determined using the coordinates given in Table 1. This method provides sufficient structural form for the unknown crystal more efficiently than attempting to determine such information *ab initio*. In addition, this method can be used to determine whether or not a given polyketide synthase in question falls within the scope of this invention.

As further disclosed herein, polyketide synthases and mutants thereof may be crystallized in the presence or absence of substrates and substrate analogs. The crystal structures of a series of complexes may then be solved by molecular replacement and compared to that of the wild-type to assist in determination of suitable replacements

for R-groups within the active site, thus making synthase mutants according to the present invention.

All mutants of the present inventions may be modeled using the information disclosed herein without necessarily having to crystallize and solve the structure for each and every mutant. For example, one skilled in the art may use one of several specialized computer programs to assist in the process of designing synthases having mutated active sites relative to the wild-type. Examples of such programs include: GRID (Goodford, 1985, J. Mod. Chem., 28:849-857), MCSS (Miranker and Karplus, 1991, Proteins: Structure, Function and Genetics, 11:29-34); AUTODOCK (Goodsell and Olsen, 1990, Proteins. Structure, Fumtion, and Genetics, 8:195-202); and DOCK (Kuntz et al., 1982, J. Mot BioL, 161:269-288), and the like, as well as those discussed in the Examples below. In addition, specific computer programs are also available to evaluate specific substrate-active site interactions and the deformation energies and electrostatic interactions resulting therefrom. MODELLER is a computer program often used for homology or comparative modeling of the three-dimensional structure of a protein. A. Saii & T.L. Blundell. J. Mol.Biol. 234:779-815, 1993. A sequence to be modeled is aligned with one or more known related structures and the MODELLER program is used to calculate a full-atom model, based on optimum satisfaction of spatial restraints. Such restraints can include, inter alia, homologous structures, site-directed mutagenesis, fluorescence spectroscopy, NMR experiments, or atom-atom potentials of mean force.

The present invention enables polyketide synthase mutants to be made and the crystal structure thereof to be solved. Moreover, by virtue of the present invention, the location of the active site and the interface of substrate therewith permit the identification of desirable R-groups for mutagenesis.

The three-dimensional coordinates of the polyketide synthase provided herein may additionally be used to predict the activity and or substrate specificity of a protein whose primary amino acid sequence suggests that it may have polyketide synthase activity. The family of CHS-related enzymes is defined, in part, by the presence of

four highly conserved amino acid residues, Cys_{164}, Phe_{215}, His_{303}, and Asn_{336}. More than 150 enzymes having these conserved residues have been identified to date, including several bacterial proteins. The functions, substrates, and products of many of these enzymes remains unknown. However, by employing the three-dimensional coordinates disclosed herein and computer modeling programs, structural comparisons of CHS can be made with a putative enzyme. Differences between the two would provide the skilled artisan with information regarding the activity and/or substrate specificity of the putative enzyme. This procedure is demonstrated in the Examples section below.

Thus, in another embodiment of the invention, there is provided a method of predicting the activity and/or substrate specificity of a putative polyketide synthase comprising (a) generating a three-dimentional representation of a known polyketide synthase using three-dimentional coordinate data, (b) generating a predicted three-dimentional representation of a putative polyketide synthase, and (c) comparing the representation of the known polyketide synthase with the representation of the putative polyketide synthase, wherein the differences between the two representations are predictive of activity and/or substrate specificity of the putative polyketide synthase.

In a further embodiment of the present invention, there is also provided a method of identifying a potential substrate of a polyketide synthase comprising (a) defining the active site of the polyketide synthase based on the atomic coordinates of said polyketide synthase, (b) identifying a potential substrate that fits the defined active site, and (c) contacting the polyketide synthase with the potential substrate of (b) and determining the activity thereon. Techniques for computer modeling and structural comparisons similar to those described herein for predicting putative polyketide synthase activity and/or substrate specificity can be used to identify novel substrates for polyketide synthases.

In addition, the structural coordinates and three-dimensional models disclosed herein can be used to design or identify polyketide synthase inhibitors. Using the

modeling techniques disclosed herein, potential inhibitor structures can be modeled with the polyketide synthase active site and those that appear to interact therewith can subsequently be tested in activity assays in the presence of substrate.

Methods of using crystal structure data to design binding agents or substrates are known in the art. Thus, the crystal structure data provided herein can be used in the design of new or improved inhibitors, substrates or binding agents. For example, the synthase polypeptide coordinates can be superimposed onto other available coordinates of similar enzymes to identify modifications in the active sites of the enzymes to create novel byproducts of enzymatic activity or to modulate polyketide synthesis. Alternatively, the synthase polypeptide coordinates can be superimposed onto other available coordinates of similar enzymes which have substrates or inhibitors bound to them to give an approximation of the way these and related substrates or inhibitors might bind to a synthase. Alternatively, computer programs employed in the practice of rational drug design can be used to identify compounds that reproduce interaction characteristics similar to those found between a synthase polypeptide and a co-crystalized substrate. Furthermore, detailed knowledge of the nature of binding site interactions allows for the modification of compounds to alter or improve solubility, pharmacokinetics, *etc.* without affecting binding activity.

Computer programs are widely available that are capable of carrying out the activities necessary to design agents using the crystal structure information provided herein. Examples include, but are not limited to, the computer programs listed below:

Catalyst Databases™ - an information retrieval program accessing chemical databases such as BioByte Master File, Derwent WDI and ACD;

Catalyst/HYPO™ - generates models of compounds and hypotheses to explain variations of activity with the structure of drug candidates;

Ludi™ - fits molecules into the active site of a protein by identifying and matching complementary polar and hydrophobic groups;

Leapfrog™ - "grows" new ligands using a genetic algorithm with parameters under the control of the user.

In addition, various general purpose machines may be used with programs written in accordance with the teachings herein, or it may be more convenient to construct more specialized apparatus to perform the operations. However, preferably the embodiment is implemented in one or more computer programs executing on programmable systems each comprising at least one processor, at least one data storage system (including volatile and non-volatile memory and/or storage elements), at least one input device, and at least one output device. The program is executed on the processor to perform the functions described herein.

Each such program may be implemented in any desired computer language (including machine, assembly, high level procedural, object oriented programming languages, or the like) to communicate with a computer system. In any case, the language may be a compiled or interpreted language. The computer program will typically be stored on a storage media or device (*e.g.*, ROM, CD-ROM, or magnetic or optical media) readable by a general or special purpose programmable computer, for configuring and operating the computer when the storage media or device is read by the computer to perform the procedures described herein. The system may also be considered to be implemented as a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner to perform the functions described herein.

Embodiments of the invention include systems (*e.g.*, internet based systems), particularly computer systems which store and manipulate the coordinate and sequence information described herein. One example of a computer system **100** is illustrated in block diagram form in Figure 9. As used herein, "a computer system" refers to the hardware components, software components, and data storage components used to analyze the coordinates and sequences as set forth in Accession Nos. 1BI5, 1D6F, 1D6I, 1D6H, 1BQ6, 1CML, 1CHW, 1CGK, 1CGZ, Table 1, and Table 3. The computer system **100** typically includes a processor for processing, accessing and manipulating

the sequence data. The processor **105** can be any well-known type of central processing unit, such as, for example, the Pentium III from Intel Corporation, or similar processor from Sun, Motorola, Compaq, AMD or International Business Machines.

Typically the computer system **100** is a general purpose system that comprises the processor **105** and one or more internal data storage components **110** for storing data, and one or more data retrieving devices for retrieving the data stored on the data storage components. A skilled artisan can readily appreciate that any one of the currently available computer systems are suitable.

In one particular embodiment, the computer system **100** includes a processor **105** connected to a bus which is connected to a main memory **115** (preferably implemented as RAM) and one or more internal data storage devices **110**, such as a hard drive and/or other computer readable media having data recorded thereon. In some embodiments, the computer system **100** further includes one or more data retrieving device **118** for reading the data stored on the internal data storage devices **110**.

The data retrieving device **118** may represent, for example, a floppy disk drive, a compact disk drive, a magnetic tape drive, or a modem capable of connection to a remote data storage system (*e.g.*, via the internet) etc. In some embodiments, the internal data storage device **110** is a removable computer readable medium such as a floppy disk, a compact disk, a magnetic tape, etc. containing control logic and/or data recorded thereon. The computer system **100** may advantageously include or be programmed by appropriate software for reading the control logic and/or the data from the data storage component once inserted in the data retrieving device.

The computer system **100** includes a display **120** which is used to display output to a computer user. It should also be noted that the computer system **100** can be linked to other computer systems **125a-c** in a network or wide area network to provide centralized access to the computer system **100**.

Software for accessing and processing the coordinate and sequences described herein, (such as search tools, compare tools, and modeling tools etc.) may reside in main memory **115** during execution.

For the first time, the present invention permits the use of molecular design techniques to design, select and synthesize novel enzymes, chemical entities and compounds, including inhibitory compounds, capable of binding to a polyketide synthase polypeptide (*e.g.*, a chalcone synthase polypeptide), in whole or in part.

One approach enabled by this invention, is to use the structure coordinates as set forth in Accession Nos. 1BI5, 1D6F, 1D6I, 1D6H, 1BQ6, 1CML, 1CHW, 1CGK, 1CGZ, Table 1, and Table 3 to design new enzymes capable of synthesizing novel polyketides. For example, polyketide synthases (PKSs) generate molecular diversity in their products by utilizing different starter molecule sand by varying the final size of the polyketide chain. The structural coordinates disclosed herein allowed the elucidation of the nature by which PKSs achieve starter molecule selectivity and control polyketide chain length. By comparing the structure of chalcone synthase, which yields a tetraketide product to 2-pyrone synthases which forms a triketide product the invention demonstrated that 2-pyrone synthase maintains a smaller initiation/elongation cavity. Accordingly, generation of a chalcone synthase mutant with an active site sterically analogous to 2-pyrone synthase resulted in the synthesis of a polyketide product of a different size. As discussed more fully below, this invention allows for the strategic development and biosynthesis of more diverse polyketides and demonstrates a structural basis for control of polyketide chain length in other PKSs. In addition, the structural coordinates allow for the development of substrates or binding agents that bind to the polypeptide and alter the physical properties of the compounds in different ways, *e.g.*, solubility.

In another approach a polyketide synthase polypeptide crystal is probed with molecules composed of a variety of different chemical entities to determine optimal sites

for interaction between candidate binding molecules (*e.g.*, substrates) and the polyketide synthase (*e.g.*, chalcone synthase).

In another embodiment, an approach made possible and enabled by this invention, is to screen computationally small molecule data bases for chemical entities or compounds that can bind in whole, or in part, to a polyketide synthase polypeptide or fragment thereof. In this screening, the quality of fit of such entities or compounds to the binding site may be judged either by shape complementarity or by estimated interaction energy. Meng, E. C. *et al.*, J. Comp. Chem., 13, pp. 505-524 (1992).

Because chalcone synthase is one member of a family of polyketide synthase polypeptides, many of which have similar functional activity, many polyketide synthase polypeptides may crystallize in more than one crystal form, the structure coordinates of chalcone synthase, or portions thereof, as provided by this invention are particularly useful to solve the structure, function or activity of other crystal forms of polyketide synthase molecules. They may also be used to solve the structure of a polyketide synthase or a chalcone synthase mutant.

One method that may be employed for this purpose is molecular replacement. In this method, the unknown crystal structure, whether it is another polyketide synthase crystal form, a polyketide synthase or chalcone synthase mutant, or a polyketide synthase complexed with a substrate or other molecule, or the crystal of some other protein with significant amino acid sequence homology to any polyketide synthase polypeptide, may be determined using the structure coordinates as provided in Accession Nos. 1BI5, 1D6F, 1D6I, 1D6H, 1BQ6, 1CML, 1CHW, 1CGK, 1CGZ, Table 1, or Table 3. This method will provide an accurate structural form for the unknown crystal more quickly and efficiently than attempting to determine such information *ab initio*.

In addition, in accordance with the present invention, a polyketide synthase or chalcone synthase polypeptide mutant may be crystallized in association or complex with known polyketide synthase binding agents, substrates, or inhibitors. The crystal

structures of a series of such complexes may then be solved by molecular replacement and compared with that of wild-type polyketide synthase molecules. Potential sites for modification within the synthase molecule may thus be identified. This information provides an additional tool for determining the most efficient binding interactions between a polyketide synthase and a chemical entity, substrate or compound.

All of the complexes referred to above may be studied using well-known X-ray diffraction techniques and may be refined to 2-3 Å resolution X-ray data to an R value of about 0.20 or less using computer software, such as X-PLOR (Yale University, 1992, distributed by Molecular Simulations, Inc.). See, *e.g.*, Blundel & Johnson, *supra*; Methods in Enzymology, vol. 114 and 115, H. W. Wyckoff *et al.*, eds., Academic Press (1985). This information may thus be used to optimize known classes of polyketide synthase substrates or binding agents (*e.g.*, inhibitors), and to design and synthesize novel classes of polyketide synthases, substrates, and binding agents (*e.g.*, inhibitors).

The design of substrates, compounds or binding agents that bind to or inhibit a polyketide synthase polypeptide according to the invention generally involves consideration of two factors. First, the substrate, compound or binding agent must be capable of physically and structurally associating with a polyketide synthase molecule. Non-covalent molecular interactions important in the association of a polyketide synthase with a substrate include hydrogen bonding, van der Waals and hydrophobic interactions, and the like.

Second, the substrate, compound or binding agent must be able to assume a conformation that allows it to associate with a polyketide synthase molecule. Although certain portions of the substrate, compound or binding agent will not directly participate in this association, those portions may still influence the overall conformation of the molecule. This, in turn, may have a significant impact on potency. Such conformational requirements include the overall three-dimensional structure and orientation of the chemical entity or compound in relation to all or a portion of the binding site, *e.g.*, active site or accessory binding site of a polyketide synthase (*e.g.*, a chalcone synthase

polypeptide), or the spacing between functional groups of a substrate or compound comprising several chemical entities that directly interact with a polyketide synthase.

The potential binding effect of a substrate or chemical compound on a polyketide synthase or the activity a newly synthesized or mutated polyketide synthase might have on a known substrate may be analyzed prior to its actual synthesis and testing by the use of computer modeling techniques. For example, if the theoretical structure of the given substrate or compound suggests insufficient interaction and association between it and a polyketide synthase, synthesis and testing of the compound may be obviated. However, if computer modeling indicates a strong interaction, the molecule may then be tested for its ability to bind to, initiate catalysis or elongation of a polyketide by a polyketide synthase. Methods of assaying for polyketide synthase activity are known in the art (as identified and discussed herein). Methods for assaying the effect of a newly created polyketide synthase or a potential substrate or binding agent can be performed in the presence of a known binding agent or polyketide synthase. For example, the effect of the potential binding agent can be assayed by measuring the ability of the potential binding agent to compete with a known substrate.

A mutagenized synthase, novel synthase, substrate or other binding compound of an polyketide synthase may be computationally evaluated and designed by means of a series of steps in which chemical entities or fragments are screened and selected for their ability to associate with the individual binding pockets or other areas of the polyketide synthase.

One skilled in the art may use one of several methods to screen chemical entities or fragments for their ability to associate with a polyketide synthase and more particularly with the individual binding pockets of a chalcone synthase polypeptide. This process may begin by visual inspection of, for example, the active site on the computer screen based on the coordinates in Accession Nos. 1BI5, 1D6F, 1D6I, 1D6H, 1BQ6, 1CML, 1CHW, 1CGK, 1CGZ, Table 1, or Table 3. Selected fragments or substrates or chemical entities may then be positioned in a variety of orientations, or docked, within an individual binding pocket of a polyketide synthase. Docking may be

accomplished using software such as Quanta and Sybyl, followed by energy minimization and molecular dynamics with standard molecular mechanics forcefields, such as CHARMM and AMBER.

Specialized computer programs may also assist in the process of selecting fragments or chemical entities. These include:

1. GRID (Goodford, P. J., "A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules", J. Med. Chem., 28, pp. 849-857 (1985)). GRID is available from Oxford University, Oxford, UK.

2. MCSS (Miranker, A. and M. Karplus, "Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method." Proteins: Structure. Function and Genetics, 11, pp. 29-34 (1991)). MCSS is available from Molecular Simulations, Burlington, Mass.

3. AUTODOCK (Goodsell, D. S. and A. J. Olsen, "Automated Docking of Substrates to Proteins by Simulated Annealing", Proteins: Structure. Function, and Genetics, 8, pp. 195-202 (1990)). AUTODOCK is available from Scripps Research Institute, La Jolla, Calif.

4. DOCK (Kuntz, I. D. *et al.*, "A Geometric Approach to Macromolecule-Ligand Interactions", J. Mol. Biol., 161, pp. 269-288 (1982)). DOCK is available from University of California, San Francisco, Calif.

Once suitable substrates, chemical entities or fragments have been selected, they can be assembled into a single polypeptide, compound or binding agent (*e.g.*, an inhibitor). Assembly may be performed by visual inspection of the relationship of the fragments to each other on the three-dimensional image displayed on a computer screen in relation to the structure coordinates of the molecules as set forth in Accession Nos. 1BI5, 1D6F, 1D6I, 1D6H, 1BQ6, 1CML, 1CHW, 1CGK, 1CGZ, Table 1, or Table 3.

This would be followed by manual model building using software such as Quanta or Sybyl.

Useful programs to aid one of skill in the art in connecting the individual chemical entities or fragments include:

1. CAVEAT (Bartlett, P. A. et al, "CAVEAT: A Program to Facilitate the Structure-Derived Design of Biologically Active Molecules". In "Molecular Recognition in Chemical and Biological Problems", Special Pub., Royal Chem. Soc., 78, pp. 182-196 (1989)). CAVEAT is available from the University of California, Berkeley, Calif.

2. 3D Database systems such as MACCS-3D (MDL Information Systems, San Leandro, Calif.). This area is reviewed in Martin, Y. C., "3D Database Searching in Drug Design", J. Med. Chem., 35, pp. 2145-2154 (1992)).

3. HOOK (available from Molecular Simulations, Burlington, Mass.).

In addition to the method of building or identifying novel enzymes or a polyketide synthase substrate or binding agent in a step-wise fashion one fragment or chemical entity at a time as described above, substrates, inhibitors or other polyketide synthase interactions may be designed as a whole or "de novo" using either an empty active site or optionally including some portion(s) of known substrates, binding agents or inhibitors. These methods include:

1. LUDI (Bohm, H.-J., "The Computer Program LUDI: A New Method for the De Novo Design of Enzyme Inhibitors", J. Comp. Aid. Molec. Design, 6, pp. 61-78 (1992)). LUDI is available from Biosym Technologies, San Diego, Calif.

2. LEGEND (Nishibata, Y. and A. Itai, Tetrahedron, 47, p. 8985 (1991)). LEGEND is available from Molecular Simulations, Burlington, Mass.

3. LeapFrog (available from Tripos Associates, St. Louis, Mo.).

Other molecular modeling techniques may also be employed in accordance with this invention. See, *e.g.*, Cohen, N. C. *et al.*, "Molecular Modeling Software and Methods for Medicinal Chemistry", J. Med. Chem., 33, pp. 883-894 (1990). See also, Navia, M. A. and M. A. Murcko, "The Use of Structural Information in Drug Design", Current Opinions in Structural Biology, 2, pp. 202-210 (1992).

Once a substrate, compound or binding agent has been designed or selected by the above methods, the efficiency with which that substrate, compound or binding agent may bind to a polyketide synthase may be tested and optimized by computational evaluation.

A substrate or compound designed or selected as a polyketide binding agent may be further computationally optimized so that in its bound state it would preferably lack repulsive electrostatic interaction with the target site. Such non-complementary (*e.g.*, electrostatic) interactions include repulsive charge-charge, dipole-dipole and charge-dipole interactions. Specifically, the sum of all electrostatic interactions between the binding agent and the polyketide synthase when the binding agent is bound to the polyketide synthase, preferably make a neutral or favorable contribution to the enthalpy of binding.

Specific computer software is available in the art to evaluate compound deformation energy and electrostatic interaction. Examples of programs designed for such uses include: Gaussian 92, revision C (M. J. Frisch, Gaussian, Inc., Pittsburgh, Pa., 1992); AMBER, version 4.0 (P. A. Kollman, University of California at San Francisco, 1994); QUANTA/CHARMM (Molecular Simulations, Inc., Burlington, Mass. 1994); and Insight II/Discover (Biosysm Technologies Inc., San Diego, Calif., 1994). These programs may be implemented, for example, using a Silicon Graphics workstation, IRIS 4D/35 or IBM RISC/6000 workstation model 550. Other hardware systems and software packages will be known to those skilled in the art of which the speed and capacity are continually modified

Once a polyketide synthase, polyketide synthase substrate or polyketide synthase binding agent has been selected or designed, as described above, substitutions may then be made in some of its atoms or side groups in order to improve or modify its binding properties. Generally, initial substitutions are conservative, *e.g.*, the replacement group will have approximately the same size, shape, hydrophobicity and charge as the original group. Such substituted chemical compounds may then be analyzed for efficiency of fit to a polyketide synthase substrate or fit of a modifed substrate to a polyketide synthase having a structure defined by the coordinates in Accession Nos. 1BI5, 1D6F, 1D6I, 1D6H, 1BQ6, 1CML, 1CHW, 1CGK, 1CGZ, Table 1, or Table 3, by the same computer methods described, above.

Conserved regions of the polyketide family synthases lend themselves to the methods and compositions of the invention. For example, pyrone synthase and chalcone synthase have conserved residues present within their active sites (as described more fully below). Accordingly, modification to the active site of chalcone synthase or a chalcone synthase substrate can be extrapolated to other conserved members of the polyketide family of synthases such as, for example, pyrone synthase.

Functional fragments of polyketide synthase polypeptides such as, for example, fragments of chalcone synthase can be designed based on the crystal structure and atomic coordinates described herein. Fragments of a chalcone synthase polypeptide and the fragment's corresponding atomic coordinates can be used in the modeling described herein. In addition, such fragments may be used to design novel substrates or modified active sites to create new diverse polyketides.

In one embodiment of the present invention, the crystal structure and atomic coordinates allow for the design of novel polyketide synthases and novel polyketide synthase substrates. The development of new polyketide synthases will lead to the development a biodiverse repetoir of polyketides for use as antibiotics, anti-cancer agents, anti-fungal agents and other therapeutic agents as described herein or known in the art. In vitro assay systems for production and determination of activity are

known in the art. For example, antibiotic activities of novel polyketides can be measured by any number of anti-microbial techniques currently used in hospitals and laboratories. In addition, anticancer activity can be determined by contacting cells having a cell proliferative disorder with a newly synthesized polyketide and measuring the proliferation or apoptosis of the cells before and after contact with the polyketide. Specific examples of apoptosis assays are provided in the following references: Lymphocyte: C. J. Li *et al.*, Science, 268:429-431, 1995; D. Gibellini *et al.*, Br. J. Haematol. 89:24-33, 1995; S. J. Martin *et al.*, J. Immunol. 152:330-42, 1994; C. Terai *et al.*, J. Clin Invest. 87:1710-5, 1991; J. Dhein *et al.*, Nature 373:438-441, 1995; P. D. Katsikis *et al.*, J. Exp. Med. 1815:2029-2036, 1995; Michael O. Westendorp *et al.*, Nature 375:497, 1995; DeRossi *et al.*, Virology 198:234-44, 1994. Fibroblasts: H. Vossbeck *et al.*, Int. J. Cancer 61:92-97, 1995; S. Goruppi *et al.*, Oncogene 9:1537-44, 1994; A. Fernandez *et al.*, Oncogene 9:2009-17, 1994; E. A. Harrington *et al.*, Embo J. 13:3286-3295, 1994; N. Itoh *et al.*, J. Biol. Chem. 268:10932-7, 1993. Neuronal Cells: G. Melino *et al.*, Mol. Cell. Biol. 14:6584-6596, 1994; D. M. Rosenbaum *et al.*, Ann. Neurol. 36:864-870, 1994; N. Sato *et al.*, J. Neurobiol 25:1227-1234, 1994; G. Ferrari *et al.*, J. Neurosci. 1516:2857-2866, 1995; A. K. Talley *et al.*, Mol. Cell Biol. 1585:2359-2366, 1995; A. K. Talley *et al.*, Mol. and Cell. Biol. 15:2359-2366, 1995; G. Walkinshaw *et al.*, J. Clin. Invest. 95:2458-2464, 1995. Insect Cells: R. J. Clem *et al.*, Science 254:1388-90, 1991; N. E. Crook *et al.*, J. Virol. 67:2168-74, 1993; S. Rabizadeh *et al.*, J. Neurochem. 61:2318-21, 1993; M. J. Birnbaum *et al.*, J. Virol 68:2521-8, 1994; R. J. Clem *et al.*, Mol. Cell. Biol. 14:5212-5222, (1994). Other assays are well within the ability of those of skill in the art.

Product of novel polyketides or polyketide synthases can be carried out in culture. For example, mammalian expression constructs carrying polyketide synthases can be introduced into various cell lines such as CHO, 3T3, HL60, Rat-1, or Jurkart cells, for example. In addition, SF21 insect cells may be used in which case the polyketide synthase gene is expressed using an insect heat shock promotor.

In another embodiment of the present invention, once a novel substrate or binding agent is developed by the computer methodology discussed above, the invention provides a method for determining the ability of the substrate or agent to be acted upon by a polyketide synthase. The method includes contacting components comprising the substrate or agent and a polyketide synthase polypeptide, or a recombinant cell expressing a polyketide synthase polypeptide, under conditions sufficient to allow the substrate or agent to interact and determining the affect of the agent on the activity of the polypeptide. The term "affect", as used herein, encompasses any means by which protein activity can be modulated, and includes measuring the interaction of the agent with the polyketide synthase molecule by physical means including, for example, fluorescence detection of the binding of an agent to the polypeptide. Such agents can include, for example, polypeptides, peptidomimetics, chemical compounds, small molecules, substrates and biologic agents as described herein. Examples of small molecules include but are not limited to small peptides or peptide-like molecules.

Contacting or incubating includes conditions which allow contact between the test agent or substrate and a polyketide synthase or modified polyketide synthase polypeptide or a cell expressing a polyketide synthase or modified polyketide synthase polypeptide. Contacting includes in solution and in solid phase. The substrate or test agent may optionally be a combinatorial library for screening a plurality of substrates or test agents. Agents identified in the method of the invention can be further evaluated by chromatography, cloning, sequencing, and the like.

Although methods and materials similar or equivalent to those described herein can be used to practice the invention, suitable methods and materials are described below. All publications, patent applications, patents and other references mentioned herein are incorporated by reference in their entirety. The invention will now be described in greater detail by reference to the following non-limiting examples.

EXAMPLES

Mutagenesis, expression, and purification. Alfalfa CHS2 cDNA (Junghans, H., et al, Plant Mol. Biol. 22:239-253, 1993) was subcloned into pHIS8 plasmid vector derived from pET-28a(+) (Novagen). PCR-based mutagenesis using the QuikChange system (Stratagene) generated the various mutants including $C_{164}S$, $C_{164}D$, $H_{303}A$, $H_{303}Q$, $H_{303}D$, $H_{303}T$, $N_{336}A$, $N_{336}D$, $N_{336}Q$, $N_{336}H$, $F_{215}S$, $F_{215}Y$ and $F_{215}W$. N-teminal His8-tagged CHS was expressed in BL21(DE3) *E. coli* cells. Cells were harvested and lysed by sonication. His-tagged CHS was purified from bacterial sonicates using a NI-NTA (Qiagen) column. Thrombin digest removed the His-tag and the protein was passed over another NI-NTA column and a benzamidine-Sepharose (Pharmacia) column. The final purification step used a Superdex 200 16/60 (Pharmacia) column.

Crystallization. CHS crystals (wild-type and $C_{164}S$ mutant) were grown by vapor diffusion at 4° C in 2 µl drops containing a 1:1 mixture of 25 mg/ml protein and crystallization buffer (2.2-2.4 M ammonium sulfate and 0.1 M PIPES, pH 6.5) in the presence or absence of 5 mM DTT. Prior to freezing at 105° K, crystals were stabilized in 40% (v/v) PEG400, 0.1 M PIPES (pH 6.5), and 0.050-0.075 M ammonium sulfate. This cryoprotectant was used for heavy atom soaks. Likewise, all substrate and product analog complexes were obtained by soaking crystals in cryoprotectant containing 10-20 mM of the compound.

Data Collection and Processing. X-ray diffraction data were collected at 105° K using a DIP2000 imaging plate system (Mac-Science Corporation, Japan) and CuK radiation produced by a rotating anode operated at 45 kV and 100 mA and equipped with double focusing Pt/Ni coated mirrors. Native CHS crystals belong to space group $P3_221$ with unit cell dimensions of $a = b = 97.54$ Å; c = 65.52 Å with a single monomer per asymmetric unit. Data were indexed and integrated using DENZO (Otwinowski & Minor, Meth. Enzymol. 276:307-326, 1997) and scaled with SCALEPACK (Otwinowski & Minor, Meth. Enzymol. 276:307-326, 1997). The

heavy atom derivative datasets were scaled against the native dataset with SCALEIT (CCP4 Suite: Programs for protein crystallography, Acta Crystallogr. D 50:760-763, 1994).

Structure determination. MIRAS was used to solve the structure of native CHS using native data set 1 (1.8 Å). Initial phasing was performed with derivative datasets including reflections to 2.3 Å resolution. Heavy atom positions for the $Hg(OAc)_2$ derivative were estimated by inspection of difference Patterson maps using the program XTALVIEW (McRee, J. Mol. Graph. 10:44-46, 1992) and initially refined with MLPHARE (Otwinowski, Z. in CCP4 Proc. *80-88,* Daresbury Laboratory, Warrington, UK, 1991). Heavy atom positions for the additional derivative data sets were determined by difference Fourier analysis using phases calculated from the $Hg(OAc)_2$ data set and the Hg positions. These sites were confirmed by inspection of difference Patterson maps. Final refinement of heavy atom parameters, identification of minor heavy atom binding sites, and phase-angle calculations were performed with the program SHARP (de La Fortelle, & Bricogne, Meth. Enzymol. 276:472-494, 1997). MIRAS phases were improved and extended to 1.8 Å by solvent flipping using the CCP4 program SOLOMON (Abrahams, & Leslie, Acta Crystallogr. D 52:30-42, 1996).

Model building and refinement. The program O (Jones, et al, Acta Crystallogr. D 49:148-157, 1993) was used for model building and graphical display of the molecules and electron-density maps. The experimental map for the native 1 dataset at 1.8 Å was of high quality and allowed unambiguous modeling of residues 3 to 389. The model was first refined with REFMAC (Murshudov, et al, Acta Crystallogr. D 53:240-255, 1997) and ARP (Lamzin & Wilson, Acta Crystallogr. D 49:129-147, 1993) against the native 1 dataset. This was followed by manual adjustments using $|2F_o-F_c|$ difference maps. Water molecules introduced by ARP were edited using the $|2F_o-F_c|$ and $|F_o-F_c|$ maps. A second refinement with SHELX-97 (Sheldrick & Schneider, Meth. Enzymol. 277:319-343, 1997) was then carried out against the native 2 data set to 1.56 Å resolution. Structures of CHS complexed with naringenin and resveratrol and the $C_{164}S$ mutant complexed with malonyl- and

hexanoyl-CoA were obtained using difference Fourier methods and were refined with REFMAC and ARP. All structures were checked with PROCHECK (Laskowski, et al, J. AppL Crystallogr. 26:283-291, 1993). 91.3 % of the residues in CHS are in the most favored regions of the Ramachandran plot, 8.4% in the additional allowed region, and 0.3% in the generously allowed region.

Three dimensional structure determination and description

Recombinant alfalfa CHS2 was expressed in *E. coli*, affinity purified using an N-terminal poly-His linker, and crystallized. The structure of wild-type CHS was determined using multiple isomorphous replacement supplemented with anomalous scattering (MIRAS) (Table X). The final 1.56 Å resolution apoenzyme model of CHS included 2982 protein atoms and 355 water molecules. In addition, the structures of a series of complexes were obtained by difference Fourier analysis. First, a crystal of a mutant ($C_{164}S$) was soaked with malonyl-CoA. This mutant retains limited catalytic activity, and the resulting acetyl-CoA complex yields insight on the decarboxylation reaction. The same mutant was also complexed with hexanoyl-CoA to mimic the structure of a linear polyketide-CoA reaction intermediate. Finally, two product analogs, naringenin and resveratrol (see Figure 1) were complexed with CHS to provide information on how the enzyme governs sequential addition of acetates to the coumaroyl moiety and how CHS controls the stereochemistry of the polyketide cyclization reaction. In plants, chalcone isomerase rapidly and stereospecifically converts chalcone to naringenin ((-)(2S)-5,7,4'-trihydroxyflavanone) through an additional ring closure. This reaction also occurs at a slower rate and non-stereospecifically in solution. As such, naringenin provides a suitable mimic of the CHS reaction product. Finally, since STS uses the same substrates as CHS but a different cyclization pathway for the biosynthesis of resveratrol, resveratrol was also soaked into CHS to investigate the structural features governing cyclization of the same substrates into two different products.

CHS functions as a homodimer of two 42 kDa polypeptides. The structure of CHS revealed that the enzyme forms a symmetric dimer with each monomer related

by a 2-fold crystallographic axis (See Figures 2a and 2b). The dimer interface buries approximately 1580 Å^2 with interactions occurring along a fairly flat surface. Two distinct structural features delineate the ends of this interface. First, the N-terminal helix of monomer A entwines with the corresponding helix of monomer B. Second, a tight loop containing a cis-peptide bond between Met_{137} and Pro_{138} exposes the methionine sidechain as a knob on the monomer surface. Across the interface, Met_{137} protrudes into a hole found in the surface of the adjoining monomer to form part of the cyclization pocket.

Each CHS monomer consists of two structural domains (see Figure 3). The upper domain exhibits an xBxBx pseudo-symmetric motif originally observed in thiolase from *Saccharomyces cerevisiae* (Mathieu, et al, Structure 2:797-808, 1994). The upper domains of CHS and thiolase are superimposeable with a r.m.s. deviation of 3.3 Å for 266 equivalent C-atoms. Both enzymes use a cysteine as a nucleophile and shuttle reaction intermediates via CoA molecules. However, CHS condenses a p-coumaroyl- and three malonyl-CoA molecules through an iterative series of reactions, whereas thiolase generates two acetyl-CoA molecules from acetoacetyl-CoA and free CoA. The drastic structural differences in the lower domain of CHS create a larger active site than that of thiolase and provide space for the polyketide reaction intermediates required for chalcone formation.

The CHS homodimer contains two functionally independent active sites. Consistent with this information, bound CoA thioesters and product analogs occupy both active sites of the homodimer in the CHS complex structures. These structures identify the location of the active site at the cleft between the upper and lower domains of each monomer. Each active site consists almost entirely of residues from a single monomer with Met_{137} from the adjoining monomer being the only exception. There are remarkably few chemically reactive residues in the active site. Four residues conserved in all the known CHS-related enzymes (Cys_{164}, Phe_{215}, His_{303}, and Asn_{336}) define the active site. Cys_{164} apparently serves as the nucleophile and as the attachment site for polyketide intermediates as previously suggested for both CHS and STS (Lanz, et al, J. Biol. Chem. 266:9971-9976, 1991). His_{303} most likely acts as a

general base during the generation of a nucleophilic thiolate anion from Cys_{164}, since the N of His_{303} is within hydrogen bonding distance of the sulfur of Cys_{164}. Phe_{215} and Asn_{336} may function in the decarboxylation reaction, as discussed below. Topologically, three interconnected cavities intersect with these four residues and form the active site architecture of CHS. These cavities include a CoA-binding tunnel, a coumaroyl-binding pocket, and a cyclization pocket.

The CoA-binding tunnel is 16 angstroms long and links the surrounding solvent with the buried active site. Binding of the CoA moiety in this tunnel positions substrates at the active site, as observed in the $C_{164}S$ mutant (described in greater detail below) complexed with malonyl- or hexanoyl-CoA. The conformation of the CoA molecules bound to CHS resembles that observed in other CoA binding enzymes. The adenosine nucleoside is in the 2'-endo conformation with an anti-glycosidic bond torsion angle. At the tunnel entrance, Lys_{55}, Arg_{58}, and Lys_{62} hydrogen bond with two phosphates of CoA. Apart from these interactions, and an additional hydrogen bond between the backbone amide nitrogen of Ala_{308} and the first carbonyl of the pantetheine moiety, van der Waals contacts dominate the remaining interactions between CHS and CoA. The pantetheine arm of the CoA extends into the enzyme positioning the terminally bound thioester-linked substrates near Cys_{164}.

Both naringenin and resveratrol bind at the active site end of the CoA-binding tunnel. The interactions observed in the naringenin and resveratrol complexes define the coumaroyl-binding and cyclization pockets (see Figure 5). The space to the lower left of the CoA-binding tunnel's end serves as the coumaroyl-binding pocket. Residues of this pocket (Ser_{133}, Glu_{192}, Thr_{194}, Thr_{197}, and Ser_{338}) surround the coumaroyl-derived portion of the bound naringenin and resveratrol molecules and interact primarily through van der Waals contacts. However, the carbonyl oxygen of Gly_{216} hydrogen bonds to the phenolic oxygen of both naringenin and resveratrol and the hydroxyl of Thr_{197} interacts with the carbonyl of naringenin derived from coumaroyl-CoA. The identity of the residues in this pocket likely contributes to the preference for coumaroyl-CoA as a substrate for parsley CHS over other cinnamoyl-CoA starter molecules, like caffeoyl- or feruloyl-CoA.

In both the naringenin and resveratrol complexes, the malonyl-derived portion of each molecule occupies a large pocket adjacent to Cys164 suggesting this is where the polyketide reaction intermediate cyclizes into the new ring system and where aromatization of the ring occurs. The six-carbon chain of hexanoyl-CoA also binds in this pocket. Physically, the size of the pocket limits the number of acetate additions to three. Phe_{265} separates the coumaroyl-binding site from the cyclization pocket and may function as a mobile steric gate during successive rounds of polyketide elongation. Although a polyketide possesses a number of hydrogen bond acceptors through which specific interactions could aid in proper folding for the cyclization reaction, the residues of the cyclization pocket, including Thr_{132}, Met_{137}, Phe_{215}, Ile_{254}, Gly_{256}, Phe_{265}, and Pro_{375}, provide few potential hydrogen bond donors. As in the coumaroyl-binding pocket, van der Waals contacts dominate the interaction between CHS and both naringenin and resveratrol. Thus, the surface topology of the cyclization pocket dictates how the malonyl-derived portion of the polyketide is folded and how the stereochemistry of the cyclization reaction leading to chalcone formation in CHS and resveratrol formation in STS is controlled.

Reaction mechanism

The position of the CoA thioesters and product analogs in the CHS active site suggest binding modes for substrates and intermediates in the polyketide elongation mechanism that are consistent with the known product specificity of CHS. In addition, the stereochemical features of the substrate and product analog complexes elucidate the roles of Cys_{164}, Phe_{215}, His_{303}, and Asn_{336} in the reaction mechanism. Utilizing structural constraints derived from the available complexes, the following reaction sequence is proposed (see Figure 6).

In the mechanism, binding of p-coumaroyl-CoA initiates the CHS reaction. Functional and structural evidence supports a coumaroyl-first mechanism over a malonyl-first one. Cerulenin, a potent irreversible inhibitor of CHS, covalently modifies Cys_{164} in CHS (Lanz, et al., J. Biol. Chem. 266:9971-9976, 1991). Preincubation of CHS with coumaroyl-CoA prevents inactivation by cerulenin, but

pre-incubation with malonyl-CoA does not (Preisig-Mueller, et al., Biochemistry 36:8349-8358, 1997). Also, the location of the coumaroyl-derived portion of naringenin and resveratrol in the CHS complexes agrees with a coumaroyl first mechanism, since the presence of a triketide reaction intermediate attached to Cys_{164} would limit access to the coumaroyl-binding pocket.

After p-coumaroyl-CoA binds to CHS, Cys_{164}, activated by His_{303}, attacks the thioester linkage, transferring the coumaroyl moiety to Cys_{164} (Monoketide Intermediate). Asn_{336} hydrogen bonds with the carbonyl oxygen of the thioester further stabilizing formation of the tetrahedral reaction intermediate. CoA then dissociates from the enzyme, leaving a coumaroyl-thioester at Cys_{164}. Binding of the first malonyl-CoA positions the bridging methylene carbon of the malonyl moiety near the carbonyl carbon of the covalently attached coumaroyl-thioester. Decarboxylation of malonyl-CoA leads to carbanion formation. Resonance between the keto and enol species stabilizes the carbanion. Attack of this carbanion on the coumaroyl-thioester releases the thiolate anion of Cys_{164} and transfers the coumaroyl group to the acetyl moiety of the CoA thioester (Diketide CoA Thioester). Capture of this elongated diketide-CoA by Cys_{164} and release of CoA sets the stage for two additional rounds of elongation resulting in formation of the tetraketide reaction intermediate.

Asn_{336} appears to play a crucial role in the decarboxylation reaction. Structural evidence shows that the decarboxylation reaction does not require transfer of the malonyl moiety to Cys_{164} as originally indicated by $C0_2$ exchange assays. Decarboxylation occurs without Cys_{164}, since the $C_{164}S$ mutant produces acetyl-CoA as determined crystallographically and confirmed by a functional assay. In the hexanoyl-CoA complex, the side chain amide of Asn_{336} provides a hydrogen bond to the carbonyl oxygen of the thioester. This interaction would stabilize the enolate anion resulting from decarboxylation of malonyl-CoA (see Figure 6). At the same time, the lack of formal positive charge at Asn_{336} may preserve the partial carbanion character of this resonance-stabilized anion, and thus the nucleophilicity of the carbanion form.

The role of Phe_{215} in the catalytic mechanism is subtler than that of Asn_{336}. Its position in both CoA complexes suggests that it provide van der Waals interactions for substrate binding. However, its conservation in bacterial enzymes related to CHS that do not make flavonoids or stilbenes may indicate a more general catalytic role for Phe_{215}. Its position near the acetyl moiety of the malonyl-CoA complex suggests that it participates in decarboxylation by favoring conversion of the negatively charged carboxyl group to a neutral carbon dioxide molecule.

Figure 7A depicts the addition of the third malonyl-CoA molecule as a three-dimensional model. The position of the coumaroyl ring in the modeled triketide intermediate is as observed in the naringenin and resveratrol complexes. The coumaroyl-binding pocket locks this moiety in position, while the acetate units added in subsequent chain extension steps bend to fill the cyclization pocket. The backbone of bound hexanoyl-CoA provides a guide for modeling the triketide reaction intermediate attached to Cys_{164}. Based on the observed acetyl-CoA complex, a rotation of the acetyl group would place the terminal methylene of the decarboxylated malonyl-CoA in position for nucleophilic attack on the triketide thioester linkage resulting in formation of a tetraketide CoA thioester.

The cyclization reaction catalyzed by CHS is an intramolecular Claisen condensation encompassing the three acetate units derived from three malonyl-CoAs. During cyclization, the nucleophilic methylene group nearest the coumaroyl moiety attacks the carbonyl carbon of the thioester linked to Cys_{164}. Ring closure proceeds through an internal proton transfer from the nucleophilic carbon to the carbonyl oxygen. Modeling of the tetraketide intermediate in a conformation leading to chalcone formation places one of the acidic protons of the nucleophilic carbon (C6) proximal to the target carbonyl (C1) (see Figure 7B). Since there is no base capable of proton abstraction from the tetraketide, it is proposed that the intermediate itself provides the driving force for carbanion formation. Protonation of the carbonyl oxygen would also stabilize the negative charge on the tetrahedral intermediate. Breakdown of this tetrahedral intermediate expels the newly cyclized ring system from Cys_{164}. Subsequent aromatization of the trione ring through a second series of

facile internal proton transfers yields chalcone.

Although the cyclization reaction has been modeled as occurring via a polyketide intermediate attached to Cys_{164}, it is possible that the reaction proceeds when the polyketide is attached to CoA. The rate of cyclization versus the rate of reattachment to Cys_{164} would dictate which of the two cyclization alternatives is mechanistically preferred.

An important question in the biosynthesis of chalcones concerns the exchangeability of the polyketide reaction intermediates. In the presence of chalcone reductase (CHR), CHS produces 6-deoxychalcone (Welle & Grisebach, FEBS Lett. 236:22-225, 1988). Mechanistically, CHR must reduce a ketone on the polyketide intermediate before cyclization occurs. Based on the CHS structure, any polyketide attached to Cys_{164} would be inaccessible to CHR unless a drastic structural change occurs in CHS upon interaction with CHR. While this conformational change is possible, such a change is difficult to imagine given the buried nature of the CHS active site. This would argue for the presence of moderately exchangeable polyketide-CoA reaction intermediates. Consistent with this idea, a recently identified CHS-like enzyme from *Pinus strobus* involved in the biosynthesis of C-methylated chalcones is active only with a starter molecule that is sterically analogous to the diketide-CoA intermediate postulated to be formed after the first condensation reaction in CHS30. These results suggest that the enzymes involved in the biosynthesis of plant polyketides may require specific localization in the plant cell to allow efficient channeling of intermediates from one enzyme to another during the production of particular products.

Cyclization specificity of CHS and STS

Both CHS and STS use the same precursor molecules and reaction mechanism to create a common tetraketide intermediate. Each enzyme must then impart a different folded conformation on this intermediate to facilitate the different cyclization reactions that yield chalcone and resveratrol. Although the three-dimensional structure of STS remains unknown, determination of the CHS structure allows

speculation about the basis for the intramolecular aldol condensation and cyclization reaction catalyzed by STS. This alternate pathway involves nucleophilic attack of the methylene group (C2) nearest the thioester linkage to Cys_{164} on the carbonyl carbon (C7) of the coumaroyl moiety (see Figure 7c). Again, modeling of the tetraketide intermediate in a conformation leading to cyclization suggests an internal proton transfer mechanism. Unlike CHS, this cyclization intermediate remains covalently attached to STS. Completion of the reaction sequence requires hydrolysis from Cys_{164} and an additional decarboxylation step prior to formation of resveratrol. These extra steps may account for the lower ,product formation rates observed with STS than with CHS (Schroeder J., et al., Biochemistry 37:8417-8425, 1998). Alternatively, the cyclization reaction may use a tetraketide-CoA thioester reaction intermediate, and subsequent hydrolysis and decarboxylation in solution.

The identity of the residue or residues involved in modulating between the intramolecular Claisen condensation in CHS and the aldol condensation in STS remains equivocal. The known CHS and STS enzymes exhibit no consistent differences in the residues lining the active site, although sequence variability between the CHS and STS enzymes does occur in the solvent exposed residues of strands ß1d (residues 253 to 259) and ß2d (residues 262-268) lining the cyclization pocket (see Figures 5b and 5c). Comparison of the naringenin and resveratrol complexes provides a possible explanation for modulation of the cyclization stereochemistry.

The cyclization pocket of CHS accommodates the newly cyclized ring of naringenin more easily than that of resveratrol. Strand ß1d (residues 253 to 259) moves slightly to enlarge the cyclization pocket in the resveratrol complex compared to the naringenin complex. Two residues that consistently vary between CHS-like and STS-like enzymes, Asp_{255} and Leu_{268}, move closer together in the resveratrol complex as ßld shifts position. Sequence variations of the solvent exposed residues of strands ß1d and ß2d may determine the conformation of the tetraketide intermediate before ring formation. Therefore, alterations in the surface topology of the cyclization pocket, mediated partially by the position of strand ßld, may affect the stereochemistry of the cyclization reaction and modulate product selectivity.

Structural basis for functionally novel CHS-like enzymes

Absolute conservation of Cys_{164}, Phe_{215}, His_{303}, and Asn_{336} occurs in CHS-like sequences, including several bacterial proteins possessing very low (typically 20-30%) amino acid sequence identity. Moreover, all CHS-like proteins exhibit strong conservation of residues shaping the geometry of the active site (Pro_{138}, Gly_{163}, Gly_{167}, Leu_{214}, Asp_{217}, Gly_{262}, Pro_{304}, Gly_{305}, Gly_{306}, Gly_{335}, Gly_{374}, Pro_{375}, and Gly_{376}). Although the functions of the bacterial CHS-like proteins remain unknown, these enzymes likely form polyketides or polyketide-CoA thioesters in a manner resembling CHS. However, steric differences resulting from sequence variation in both the coumaroyl-binding pocket and the cyclization pocket strongly suggest alternate substrate and product specificity in the bacterial enzymes.

The sequence databases include approximately 150 plant enzyme sequences classified as CHSlike proteins. The substrate and product specificity of a majority of these sequences remains to be determined. In addition, the high sequence similarity of all plant sequences complicates classification of these sequences as authentic CHS, STS, ACS, or BBS enzymes. The information provided by the three-dimensional structure of CHS should make new substrate and product specificity more readily discernible from sequence information.

To illustrate the usefulness of structural information in identifying potentially new activities, a CHS-related sequence from *Gerbera hybrids* (GCHS2)32 that is 74% identical with alfalfa CHS2 was examined. Modeling the active site architecture of GCHS2 using the structure of alfalfa CHS2 as a template indicates that GCHS2 will not catalyze either the CHS-like or STS-like reaction (see Figure 8). This variation in reaction specificity results from striking steric differences in the coumaroyl binding and cyclization pockets that substantially reduce the volume of both pockets from 923 $Å^3$ in CHS to 269 $Å^3$ in GCHS2. Side chain variation at positions 197 and 338 alter the coumaroyl binding pocket, while the identity of residue 256 dictates major steric changes in the cychzation pocket. The reduced size of these pockets in GCHS2 suggests that fewer than three acetate additions will occur, and that a CoA thioester

with an acyl moiety smaller than p-coumaroyl initiates the reaction. Recent functional characterization of GCHS2 confirms this prediction and demonstrates that this enzyme uses acetyl-CoA or benzoyl-CoA and two condensation reactions with malonyl-CoA to form pyrone products (Eckermann, et al., Nature 396:397-396, 1998).

Crystallization of Additional Polyketide Synthases

Stilbene synthase from *Pinus strubus* was overexpressed in *E. coli* as an octahistidyl N-terminal fusion protein, purified to >90% homogeneity by metal affinity and gel filtration chromatography, and crystallized in the preparation lacking the N-terminal polyhistidine tag (removed by thrombin cleavage) from 13% (w/v) polyethylene glycol (PEG8000), 0.05 M MOPSO, 0.3 M ammonium acetate at pH 7.0. This STS is 396 amino acids in length and, like alfalfa CHS exists as a homodimer in solution. A partial data set on a frozen crystal (!))K) has been collected to 2.7 Å. The crystals belong to space group C222 with with unit cell dimensions of a = 74.94 Å , b = 86.63 Å , c = 364.18 Å , $\alpha = \beta = \gamma = 90°$.

2-Pyrone synthase (2-PS) from *Gerbera hybrida* was expressed and purified from *E. coli* in a similar manner to CHS and STS. Crystals were obtained from 1.5 M ammonium sulfate, 011 M Na^+ - succinate, 0.002 M DTT at pH 5.5.

2-Pyrone synthase (2-PS) from *Gerbera hybrida* forms a triketide from an acetyl-CoA initiator and two acetyl-CoA α-carbanions derived from decarboxylation of two malonyl-CoAs that cyclizes into the 6-methyl-4-hydroxy-2-pyrone. In comparison, alfalfa chalcone synthase 2 (CHS2; 74% amino acid sequence identity to 2-PS), condenses *p*-coumaroyl-CoA and three acetyl-CoA α-carbanions derived from decarboxylation of three malonyl-CoAs into a tetraketide that cyclizes into chalcone. A homology model of 2-PS based on the structure of CHS suggested that the 2-PS initiation/elongation cavity is smaller than that of CHS. A smaller cavity would account for the terminal formation of a triketide intermediate prior to cyclization by 2-PS.

Expression, Purification and Crystallization of 2-PS.

2-PS was expressed in *E. coli*, purified and crystallized as described above. *Gerbera hybrida* 2-PS was expressed in *E. coli* using the pHIS8 vector and was purified as described for CHS. 2-PS crystals grew at 4 °C in hanging-drops containing a 1:1 mixture of 25 mg ml^{-1} protein and crystallization buffer (1.5 M ammonium sulfate, 50 mM succinic acid (pH 5.5), and 5 mM DTT). Before freezing at 105 K, crystals ($P3_121$; unit cell dimensions a = 82.15 Å, c = 241.33 Å; one 2-PS dimer per asymmetric unit) were stepped through stabilizer (50 mM succinic acid (pH 5.5), 50 mM ammonium sulfate, and 5 mM DTT) containing 5 mM acetoacetyl-CoA and increasing concentrations of glycerol (30% (v/v) final). Diffraction data were collected using a DIP2030 imaging plate system and CuK radiation produced by a rotating anode (wavelength 1.54 Å). All images were processed with DENZO/SCALEPACK (Z. Otwinowski, W. Minor, *Methods Enzymol.* 276:307 (1997)). A total of 179,623 reflections were merged to give 60,824 unique reflections (98.2% complete overall to 2.05 Å and 98.1% complete in the highest resolution shell) with an R_{sym} = 0.042 (0.206 in the highest resolution shell) and an I/_of 21.7 (4.5 in the highest resolution shell). The structure of 2-PS complexed with acetoacetyl-CoA was determined by molecular replacement using CHS as a search model and was refined to 2.05 Å resolution. The overall fold of 2-PS is the αβαβα motif found in CHS and β-ketoacyl synthase II (KAS II). In addition, the positions of the catalytic residues of 2-PS (Cys_{169}, His_{308}, and Asn_{341}), CHS (Cys_{163}, His_{303}, Asn_{336}), and KAS II (Cys_{163}, His_{303}, and His_{340}) are structurally analogous. As expected from sequence homology, the structures of 2-PS and CHS are nearly identical and superimpose with a r.m.s. deviation of 0.64 Å for the two proteins' α-carbon atoms. Similar to CHS, the 2-PS dimerization surface buries 1805 $Å^2$ of surface area per monomer and a loop containing a *cis*-peptide bond between Met_{142} and Pro_{143} allows

the methionine of one monomer to protrude into the adjoining monomer's active site. Thus, dimerization allows formation of the complete 2-PS active site.

Acetoacetyl-CoA is a reaction intermediate of 2-PS. Electron density for the ligand is well defined in the 2-PS active site and shows that the acetoacetyl moiety extends from the CoA pantetheine arm into a large internal cavity. The electron density also reveals oxidation of the catalytic cysteine's (Cys_{169}) sulfhydryl to sulfinic acid (-SO_2H). This oxidation state prevents formation of a covalent acetoacetyl-enzyme complex but allows trapping of the bound acetoacetyl-CoA intermediate. Extensive protein-ligand contacts position CoA at the entrance to the active site and orient the acetoacetyl moiety at the end of a 15 Å long tunnel that opens into a cavity that defines the initiation and elongation steps of polyketide formation.

The 2-PS active site cavity consists of twenty-seven residues from one monomer and Met_{142} from the adjoining monomer. Phe_{220} and Phe_{270} mark the boundary between the CoA binding site and the initiation/elongation cavity. Near the CoA thioester, Cys_{169}, His_{308}, and Asn_{341} form the catalytic center of 2-PS. These residues are conserved in all homodimeric iterative PKSs. Based on this, catalytic roles were proposed for each residue that are analogous to the corresponding residues in CHS. Cys_{169} acts as the nucleophile in the reaction and as the attachment site for the elongating polyketide chain. Interaction between His_{308} and Cys_{169} maintains the thiolate required for condensation of the starter molecule. His_{308} and Asn_{341} catalyze malonyl-CoA decarboxylation and stabilize the transition states during the condensation steps by forming an oxyanion hole that accommodates the negatively charged tetravalent transition state. Following the first condensation reaction, a diketide remains attached to Cys_{169}. The second malonyl-CoA then binds, undergoes decarboxylation, and the resulting nucleophilic acetyl-coA α-carbanion performs a

second condensation reaction with the enzyme bound diketide, ultimately generating the triketide that cyclizes into methylpyrone.

Comparison of the initiation/elongation cavities of 2-PS and CHS reveal four amino acid differences. In 2-PS, Leu_{202}, Met_{259}, Leu_{261}, and Ile_{343} replace Thr_{197}, Ile_{254}, Gly_{256}, and Ser_{338}, respectively, of CHS. These four substitutions reduce cavity volume from 923 $Å^3$ in CHS to 274 $Å^3$ in 2-PS. A model of methylpyrone in the 2-PS cavity, based on the position of acetoacetyl-CoA, emphasizes the volume change compared to the CHS-naringenin complex (Accession No. 1CGK). Leu_{202} and Ile_{343} occlude the portion of the 2-PS cavity corresponding to the coumaroyl-binding site of CHS. Replacement of Gly_{256} in CHS by Leu_{261} in 2-PS severely reduces the size of the active site cavity. Substitution of Met_{259} in 2-PS for Ile_{254} in CHS produces a modest alteration in cavity volume. To examine the functional importance of these amino acid differences, the initiation/elongation cavity of CHS was altered by mutagenesis to resemble that of 2-PS. The resulting mutant proteins were screened for activity using either *p*-coumaroyl-CoA or acetyl-CoA as starter molecules. Activities of 2-PS, CHS, and the CHS mutants were determined by monitoring product formation using a TLC-based radiometric assay. Assay conditions were 100 mM Hepes (pH 7.0), 30 µM starter-CoA (either *p*-coumaroyl-CoA or acetyl-CoA), and 60 µM [^{14}C]-malonyl-CoA (50,000 cpm) in 100 µl at 25 °C. Reactions were quenched with 5% acetic acid, extracted with ethyl acetate, and applied to TLC plates and developed. Due to the spontaneous cyclization of chalcone into the flavanone naringenin, activities of CHS are referenced to naringenin formation.

The x-ray crystal structures of 2-PS and CHS imply that the size of the active site cavity limits polyketide length and modulates folding of the polyketide chain. Wild-type CHS generates the tetraketide chalcone and 2-PS produces the triketide methylpyrone. Likewise, the CHS I254M mutant also yields chalcone. Interestingly,

the T197L, G256L, and S338I mutants do not form chalcone. Crystallographic analysis of the G256L and S338I mutants demonstrates that the substituted side-chains adopt conformations similar to the corresponding residues in 2-PS without altering the position of the protein backbone. Since the T197L, G256L, and S338I mutants altered product formation, a CHS triple mutant was generated. Consistent with the proposal that cavity volume dictates polyketide length, the T197L/G256L/S338I mutant produces only methylpyrone, as confirmed by liquid chromatography/mass spectroscopy (LC/MS). LC/MS/MS analysis was performed by the Mass Spectroscopy facility of the Scripps Research Institute. Scaled-up assays (2 ml reaction volume) with the CHS T197L/G256L/S338I mutant and 2-PS were performed. Extracts were analyzed on a Hewlett-Packard HP1100 MSD single quadrupole mass spectrometer coupled to a Zorbax SB-C_{18} column (5 μm, 2.1 mm x 150 mm). HPLC conditions were as follows: gradient system from 0 to 100% methanol in water (each containing 0.2% acetic acid) within 10 min; flow rate 0.25 ml min^{-1}. LC/MS/MS data from both reactions were identical: 6-methyl-4-hydroxy-2-pyrone, R_t = 5.068 min; $[M-H]^-$ 125 (41); $[M-H-CO_2]^-$ 81 (100). The numbers show *m/z* values with relative intensities in parenthesis. The observed fragmentation matches previously published data.

In addition, the size of the cavity in 2-PS and CHS confers starter molecule specificity. 2-PS accepts acetyl-CoA but does not use *p*-coumaroyl-CoA. Structurally, the constricted 2-PS active site excludes the bulky coumaroyl group. As such, incubation of 2-PS in the presence of coumaroyl-CoA and malonyl-CoA yields methylpyrone produced from three malonyl-CoA molecules. In comparison, the larger initiation/elongation cavity of CHS allows for different sized aliphatic and aromatic starter molecules to be used *in vitro* with varying efficiencies. CHS exhibits a 230-fold preference for *p*-coumaroyl-CoA versus acetyl-CoA. Alterations in the

active site cavity of CHS, affect starter molecule preference. The CHS I254M mutant is functionally comparable to wild-type enzyme with a modest reduction in specific activity. The T197L and S338I mutants exhibit 10-fold and 3-fold preferences, respectively, for coumaroyl-CoA. Moreover, both form a distinct product using coumaroyl-CoA as a starter molecule. In contrast, the G256L mutant favors acetyl-CoA 3-fold. Like 2-PS, the CHS T197L/G256L/S338I (3x) mutant only accepts acetyl-CoA (or malonyl-CoA) as the starter molecule.

Functional diversity among other homodimeric iterative PKSs, like *p*-coumaroyltriacetic acid synthase (CTAS), acridone synthase (ACS), and the *rppA* protein from *Streptomyces griseus*, likely results from variations of residues lining the initiation/elongation cavity. As demonstrated, positions 197, 256, and 338 distinguish between tetraketide products derived from a final Claisen condensation in wild-type CHS and triketide products derived from an enolate-directed condensation in the CHS triple mutant. Although CHS, CTAS, and ACS generate tetraketides, each enzyme differs in either the cyclization reaction or in the identity of the starter molecule. CTAS forms the same enzyme-bound tetraketide as CHS but does not catalyze the final cyclization reaction. Comparison of these two enzymes reveals that substitution of Thr 197 in CHS with an asparagine in CTAS may prevent the covalently-bound tetraketide intermediate from undergoing cyclization into chalcone. ACS uses N-methylanthranoyl-CoA as a starting substrate to produce the alkaloid acridone. Three differences between CHS (Thr_{132}, Ser_{133}, and Phe_{265}) and ACS (Ser_{132}, Ala_{133}, and Val_{265}) may alter starter molecule specificity. In ACS, these changes likely widen the portion of the cavity corresponding to the *p*-coumaroyl-binding site in CHS to accommodate N-methylanthranoyl-CoA binding. Comparative changes in the active site cavity allow formation of longer polyketides. The *rppA* protein forms a pentaketide from five acetates derived from malonyl-CoA decarboxylation. Thr_{137}, Ala_{138}, Thr_{199}, Leu_{202}, Met_{259}, Leu_{261}, Leu_{268}, Pro_{304}, and Ile_{343} of 2-PS are replaced by

Cys_{106}, Thr_{107}, Cys_{168}, Cys_{171}, Ile_{228}, Tyr_{230}, Phe_{237}, Ala_{261}, and Ala_{295}, respectively, in the *rppA* protein. Models of the *rppA* protein based on the 2-PS and CHS structures show that cavity volume is 1145 $Å^3$ in the *rppA* protein versus 274 $Å^3$ in 2-PS (or 923 Å in CHS). Manipulation of the active site through amino acid substitutions offers a strategy for increasing the molecular diversity of polyketide formation through both the choice of starter molecule and the number of subsequent condensation steps.

The reaction mechanism for polyketide formation and the structural basis for controlling polyketide length described here may be shared with other more complex iterative (*e.g.*, actinorhodin (*act*) PKS and tetracenomycin (*tcm*) PKS) and modular PKSs (*e.g.*, 6-deoxyerythronolide B synthase (DEBS)). The structural similarity of the 2-PS, CHS, and KAS II active sites, the sequence homology of KAS II and the ketosynthases of *act* PKS, *tcm* PKS, and DEBS, and mutagenesis studies of CHS and *act* PKS demonstrating similar roles for the catalytic residues of each protein indicate that a conserved active site architecture catalyzes similar reactions in these enzymes.

As in 2-PS and CHS, the volume of the active site cavities in other PKSs likely limits the size of the final polyketide. For example, *act* PKS and *tcm* PKS generate octaketide and decaketide products, respectively, at a single active site. This suggests that the active site cavities of these PKSs differ in size, and are larger than those of 2-PS or CHS. Similarly, the ketosynthases of different DEBS modules accept polyketide intermediates ranging in length from five to twelve carbons. Modular PKSs, like DEBS, use an assembly-line system in which an individual module catalyzes one elongation reaction and passes the growing polyketide to the next module. Although the ketosynthase domains of DEBS are functionally permissive, modulation of active site volume in each module's ketosynthase would provide selectivity for the proper sized intermediate at each elongation step.

Structural differences among PKSs alter the volume of the initiation/elongation cavity to allow discrimination between starter molecules and to vary the number of elongation steps to ultimately direct the nature and length of the polyketide product.

While the foregoing has been presented with reference to particular embodiments of the invention, it will be appreciated by those skilled in the art that changes in these embodiments may be made without departing from the principles and spirit of the invention, the scope of which is defined by the appended claims.

That which is claimed:

1. An isolated polyketide synthase comprising at least fourteen active site α-carbons having the structural coordinates of Table 1.

2. The isolated polyketide synthase of claim 1, wherein the amino acid located at position 164 is alanine or serine.

3. The isolated polyketide synthase of claim 1, wherein the amino acid located at position 303 is alanine, asparagine, glutamine, aspartic acid, or threonine.

4. The isolated polyketide synthase of claim 1, wherein the amino acid located at position 336 is a lysine, alanine, aspartic acid, glutamine, or histidine.

5. The isolated polyketide synthase of claim 1, wherein the amino acid located at position 215 is serine, tyrosine, or tryptophan.

6. The isolated polyketide synthase of claim 1, wherein the polyketide synthase has atomic coordinates as set forth in PDB Accession Nos: 1BI5, 1BQ6, 1CML, 1CHW, 1CGK, 1CGZ. 1D6F, 1D6I, or 1D6H.

7. A nucleic acid encoding the synthase of claim 1.

8. A nucleic acid encoding the synthase of claim 2.

9. A nucleic acid encoding the synthase of claim 3.

10. A nucleic acid encoding the synthase of claim 4.

11. A nucleic acid encoding the synthase of claim 5.

12. A method of predicting the activity and/or substrate specificity of a putative polyketide synthase, said method comprising:

comparing the representation of a known polyketide synthase and the representation of a putative polyketide synthase, wherein differences between the two

representations are predictive of polyketide synthase activity and/or substrate specificity.

13. The method of claim 12, wherein the known polyketide synthase is chalcone synthase, stilbene synthase, or pyrone synthase.

14. The method of claim 13, wherein the known chalcone synthase has strucutral coordinants as set forth in PDB Accession Nos: 1BI5, 1BQ6, 1CML, 1CHW, 1CGK, or 1CGZ.

15. The method of claim 13, wherein the known pyrone synthase has atomic coordinates as set forth in Table 3.

16. The method of claim 12, wherein the putative synthase is a mutant of a known polyketide synthase.

17. A crystalline form of the polyketide synthase of claim 1.

18. A crystalline form of the polyketide synthase of claim 2.

19. A crystalline form of the polyketide synthase of claim 3.

20. A crystalline form of the polyketide synthase of claim 4.

21. A crystalline form of the polyketide synthase of claim 5.

22. A crystalline chalcone synthase, stilbene synthase, or pyrone synthase..

23. A crystalline complex comprising chalcone synthase and a chalcone synthase substrate.

24. The crystalline complex of claim 23, wherein the chalcone synthase is native chalcone synthase.

25. The crystalline complex of claim 23, wherein the chalcone synthase is a non-native chalcone synthase.

26. The crystalline complex of claim 23, wherein the chalcone synthase substrate is selected from the group consisting of chalcone, naringenin, resveratrol, cerulenin, acyl-CoA, malonyl-CoA, and hexanoyl-CoA.

27. The crystalline complex of claim 23, wherein the complex has atomic coordinates as set forth in PDB Accession Nos: 1BQ6, 1CML, 1CHW, 1CGK or 1CGZ.

28. A method of identifying a potential substrate of a polyketide synthase, said method comprising:

(a) defining the active site of said polyketide synthase based on a plurality of atomic coordinates of said polyketide synthase,

(b) identifying a potential substrate that fits the active site of (a) with the polyketide synthase, and

(c) contacting the polyketide synthase with the potential substrate and determining its activity thereon.

29. The method of claim 28, wherein the polyketide synthase is chalcone synthase, stilbene synthase, or pyrone synthase.

30. The method of claim 28, wherein the polyketide synthase is a mutant of a known polyketide synthase.

31. The method of claim 30, wherein the known polyketide synthase is chalcone synthase, stilbene synthase, or pyrone synthase.

32. The method of claim 28, wherein the plurality of atomic coordinates are as set forth in PDB Accession Nos: 1BI5, 1BQ6, 1CML, 1CHW, 1CGK, 1CGZ. 1D6F, 1D6I, 1D6H, or portions thereof.

33. A method of identifying a potential inhibitor of a polyketide synthase, said method comprising:

(a) defining the active site of said polyketide synthase based on a plurality of atomic coordinates of said polyketide synthase,

(b) contacting a potential compound that fits the active site of (a) with the polyketide synthase in the presence of a substrate, and

(c) determining the ability of said compound to inhibit the activity of said polyketide synthase on said substrate.

34. The method of claim 33, wherein the polyketide synthase is chalcone synthase, stilbene synthase, or pyrone synthase.

35. The method of claim 33, wherein the polyketide synthase is a mutant of a known polyketide synthase.

36. The method of claim 35, wherein the mutant polyketide synthase is a mutant of chalcone synthase, stilbene synthase, and pyrone synthase.

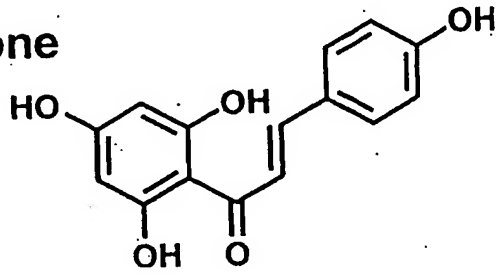
37. The method of claim 33, wherein the plurality of atomic coordinates are as set forth in PDB Accession Nos: 1BI5, 1BQ6, 1CML, 1CHW, 1CGK, 1CGZ. 1D6F, 1D6I, 1D6H, or portions thereof.

38. A computer program on a computer readable medium, said computer program comprising instructions to cause a computer to:

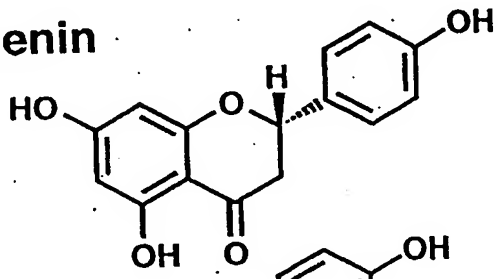
define a polyketide synthase or fragment thereof based on a plurality of atomic coordinates of the polyketide synthase.

39. The computer program of claim 38, wherein the plurality of atomic coordinates are as set forth in PDB Accession Nos: 1BI5, 1BQ6, 1CML, 1CHW, 1CGK, 1CGZ. 1D6F, 1D6I, 1D6H, Table 3, or portions thereof.

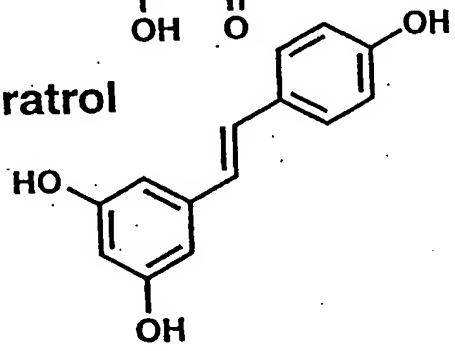
Chalcone



Naringenin



Resveratrol



Cerulenin

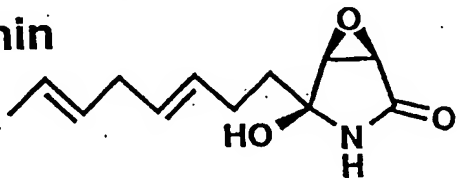


FIGURE 1A

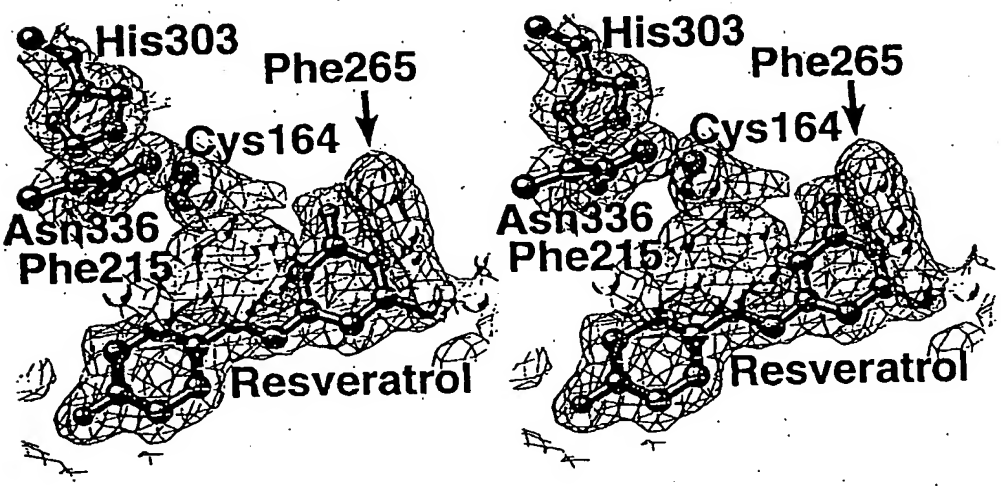

His303
Phe265
Cys164
Asn336
Phe215
Resveratrol
His303
Phe265
Cys164
Asn336
Phe215
Resveratrol

FIGURE 1B

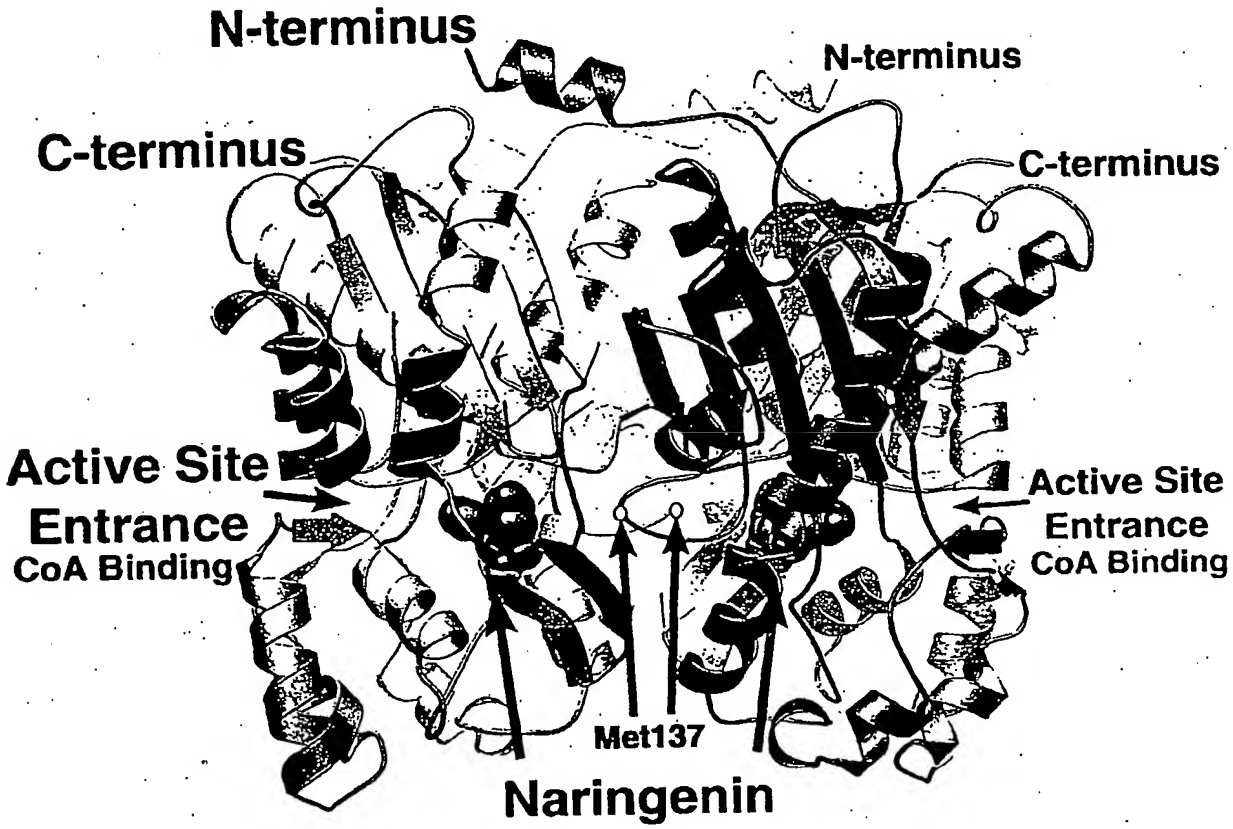

N-terminus
N-terminus
C-terminus
C-terminus
Active Site
Entrance
CoA Binding
Active Site
Entrance
CoA Binding
Met137
Naringenin

Figure 2A


3
183
283
223
243
323
63
263
83
43
3
183
283
223
243
323
63
263
83
43

FIGURE 2B